This is a preliminary version of an article published in
Proc. of ACM Multimedia Systems (MMSYS), pp. 23-34, Phoenix, Arizona, USA, February 2010
by Philipp Schaber, Stephan Kopf, Wolfgang Effelsberg, Niels Thorwirth

Semi-automatic Registration of Videos for Improved Watermark Detection

Philipp Schaber
Department of Computer Science IV
University of Mannheim
68131 Mannheim, Germany
schaber@informatik.uni-mannheim.de

Stephan Kopf
Department of Computer Science IV
University of Mannheim
68131 Mannheim, Germany
kopf@informatik.uni-mannheim.de

Wolfgang Effelsberg
Department of Computer Science IV
University of Mannheim
68131 Mannheim, Germany
effelsberg@informatik.uni-mannheim.de

Niels Thorwirth
Verimatrix Inc.
CA 92121, San Diego
nthorwirth@verimatrix.com

ABSTRACT

Virtually every video watermarking technology can benefit from comparison with the original content. For non-blind schemes it is fundamental; for others it is an improvement to increase the watermark's signal-to-noise ratio by subtracting the content that is often noise to the detector. A direct frame-by-frame comparison of the videos is not possible due to the fact that illegal copies of videos usually differ significantly from their originals caused by different spatial resolution or frame rates, geometric distortions from capturing, or targeted attacks. In this paper, we present a software tool that enables the *semi-automatic* temporal *and* spatial synchronization of frames and pixels of two similar videos. This process is called *registration*. We put our focus on utilizing human capabilities with the smallest possible effort, to allow a high overall performance and precision of the registration. An efficient graphical user interface supports the users and visualizes the results of all steps. In addition, we specifically distinguish digitally reproduced copies from analog (camcorded) copies in which two or more frames are blended into a new frame.

Categories and Subject Descriptors

I.4.3 [**Image Processing and Computer Vision**]: Enhancement—*Registration*; D.4.6 [**Security and Protection**]: Authentication

General Terms

Security, Verification, Algorithms

Keywords

Digital watermarking, spatial registration, temporal registration

MMSys'10, February 22–23, 2010, Phoenix, Arizona, USA.

1. INTRODUCTION

A major concern with the digital distribution of high-valued content such as movies is theft by piracy, and digital forensic watermarking as an anti-piracy tool has recently gained increased attention. An invisible digital watermark can be used to track copyrighted content and determine where and when illegal distribution occurred, without limiting legal use. In order to effectively track illegal distribution, watermarking schemes not only have to be robust against distortions, but must also be secure against unauthorized removal and embedding. In addition, they shall not alter the quality of the marked content. This poses significant challenges to a watermark design.

Several watermarking technologies enable the extraction of the embedded data based on the analysis of the modified video alone. Nevertheless, the reliability of the forensic analysis is much higher if the original video without the embedded watermark is also taken into account during the analysis, and some watermarking algorithms even require the availability of a reference video.

To achieve a frame and pixel-wise comparison, the original video first needs to be registered to the watermarked copy. Each copy is a modified version of the original, and depending on how it was created (e.g., recording of movies with a camcorder), significant misalignments exist between both versions. For many common video processes, a combination of *spatial* and *temporal* misalignments is introduced, as well as quality degradations. Spatial misalignment is the result of geometric distortions such as cropping or aspect ratio change, while the change of frame rate or cutting causes temporal misalignment. Another aspect of camcorded copies is the fact that frame intervals displayed are not synchronized with the recording, and thus two or more source frames are blended into a single frame of the recording.

In this paper, we present a novel system that enables the semi-automatic temporal and spatial registration of two videos. The distinct features of our approach are:

- The system supports both temporal *and* spatial alignment, whereas many system focus on one aspect only.
- We see a huge benefit to the overall registration performance in utilizing human capabilities, but our goal is

to do so with the smallest possible effort. This is why we present a *semi-automatic* approach, having a manual initialization combined with a precise automatic registration.

- The algorithms differentiate between copies resulting from analog (camcorded) and digital reproduction. Most systems do not distinguish between the two and ignore the fact that implications and assumable constraints are very different, especially for temporal distortions.

- Professional users can work very efficiently with our software tool due to the high functionality of the graphical user interface. As we present a semi-automatic approach, our implementation ensures a smooth work flow of all tasks, keeping the manual part as short as possible. To support videos up to HD resolution, multithreading utilizes the power of modern multi-core processors.

The paper is structured as follows: In the following Section, we describe some fundamental concepts of digital watermarking in videos. Section 3 describes the semi-automatic spatial and temporal video registration algorithm and the user interface of the system. Experimental results are presented in Section 4. Related work is then presented in Section 5. The paper concludes with an outlook in Section 6.

2. FUNDAMENTALS AND OVERVIEW

In this section, we give an introduction to digital watermarking, focus on typical misalignments of videos, and explain the benefits of video registration. An overview of the watermarking tool chain and the positioning of our registration system is given in Figure 1.

2.1 Introduction to Digital Watermarking

Digital watermarking is a technique of embedding additional information in host data, most often into media data such as pictures, audio or video data. Contrary to metadata, where information is stored alongside the host data, watermarks store the information in the content itself by modifying it. These modifications should be imperceptible to human observers. Nevertheless, it is possible to read the embedded information afterwards with an appropriate watermark detector.

The process of inserting a watermark signal into a host signal is called **embedding**. To ensure invisibility, the embedding usually employs a perceptual model to control the position and amount of modifications that may thus vary in different parts of the host signal. Reading out the embedded signal is most often referred to as **detection** or **extraction** of the watermark. The need for the original data during extraction categorizes watermarking schemes: With **blind extraction** watermarking does not need the unmarked, original host data to retrieve the watermark (although, it typically will profit significantly from its availability). On the other hand, **non-blind** watermarking requires the unmodified content for extraction. To increase security of a watermarking system, encryption can be used to encrypt the embedded information. A secret key is generated for embedding, which prevents extraction or modification of the watermark without knowledge of the key – even when the watermarking scheme is publicly known.

Watermarking schemes can be further distinguished regarding their intended behavior to incidental as well as hostile modifications and distortions. If a scheme is classified as **fragile**, the desired behavior of the watermark is to immediately degrade when a modification is performed, which guarantees the identification of alterations to the content. If it is classified as **robust**, it should survive distortions (e.g., rotation, scaling and translation in case of images [28]) and remain extractable even after severe degradations. The desired application scenario decides on the required scheme. Tracking pirated copies in order to identify and stem illegal distribution of movies is one of the main fields for robust watermarking today. In our scenario, we focus on *robust* digital watermarking of *videos*.

For tracking, a personalized copy needs to be created for everybody legally receiving the media content, each with a different watermarking "payload". In this way, the origins of pirated copies can be found; that makes it a good complement to digital rights management (DRM) or any other active content security scheme. A unique, traceable identifier is embedded in each individual copy as soon as the media leaves the (DRM-)protected domain. For example, a video stream might be encrypted on its distribution channel, but as soon as it gets displayed it has to be decrypted in order to present it. Client-side watermarks typically get embedded at this point. To identify the origin, a customer or transaction ID can be used or any other information that allows accurate tracking.

Besides invisibility, a high robustness is required since removing the watermark or making it undetectable is the primary goal of hostile attacks. Even if a copy is not the target of any attack, the watermark is supposed to survive all common signal processing operations and remain in the media throughout the complete (legal and illegal) distribution chain. Thus, the complete life cycle of a watermark is usually not only modeled by embedding and extraction, but consists of three phases, namely:

$$\text{Embedding} \rightarrow \text{Attack} \rightarrow \text{Extraction}$$

After the content has been watermarked, it usually gets distributed in some way. Any modification from then on is called *attack*, even though modifications do not need to be targeted. The term "attack" comes from tracking and copyright protection applications of watermarks, where intentional, hostile modifications aim to render the watermark undetectable. Other, non-intentional modifications are introduced by many common signal processing functions, and are unavoidable when the video travels down the distribution chain. All these modification result in misalignments between the original, unaltered data and a distorted copy. Figure 1 gives an overview of the watermarking life cycle and the three phases embedding, attack, and extraction.

2.2 Misalignments in Videos

For digital media, any number of copies will not alter a video and thus will not cause misalignments. However, although plain digital copies are lossless, on both analog *and* digital distribution channels media is still getting distorted since distribution not only involves copying. Deployment of media to different end-users devices requires different formats of the content itself. From the many video distribution formats that exist today, most often each is specialized in supporting especially one distribution channel.

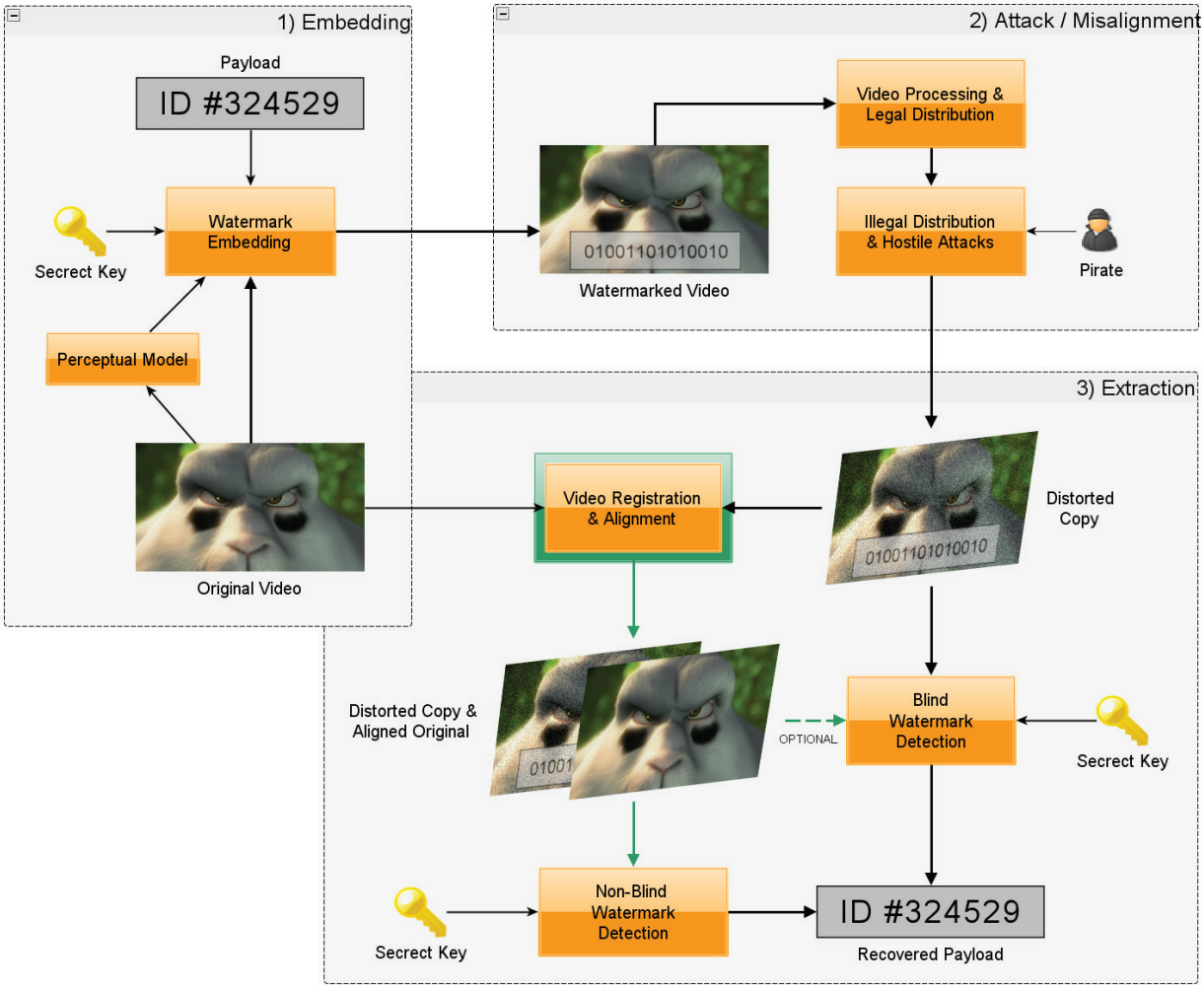


Figure 1: System overview

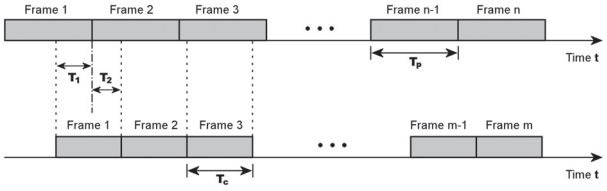
Different transformations (or a combination of these) may be applied to a video in order to convert it to a specific format, allowing a specific kind of distribution. To prepare for illegal distribution the content is typically formatted for download speed, to fit within the boundaries of a physical medium or for playback on a computer. Usual transformations include: Reducing the video's resolution (down-sampling), frame rate changes, changes from interlaced to progressive scan or vice-versa, color space conversions, aspect ratio changes with cropping or letterbox, and re-compression with different compression codecs and color samplings.

A fundamental differentiation can be made by distinguishing modifications that result from pure digital versus analog processing. Although digital video formats are already dominant in legal offline and online distribution systems, and almost all movie piracy uses digital channels, too, considering modifications from analog channels is still relevant, especially in the context of fighting piracy: Digital watermarks are capable of passing through the so-called **"analog hole"**. The re-recording of a displayed video with a camera can not be prevented with encryption, nor do any reliable and secure technologies exist to prevent this attack. And even though illegal *distribution* is mainly digital, the *acquisition* is often analog, for two reasons:

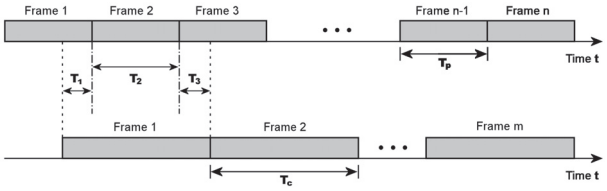
- The mentioned analog hole is used to circumvent copy-protection systems (e.g., digitally capturing video playback), or:
- There is no direct access to the medium, so "remote" techniques such as camcording a displayed video in the cinema are used, utilizing the screen as an analog channel.

Based on the previous considerations, we define the video reproduction method in the following way: We speak of *digital reproduction* if the video is processed digitally in all steps; otherwise it is an *analog* or *camcorded reproduction*.

Videos based on analog and digital reproduction usually include different spatial and temporal modifications. Considering *spatial misalignments*, the geometric distortions found after digital processing are mostly combinations of *scaling*, *cropping* and *translations* only, as resulting from format conversions. Usually, those transformations are well-defined



(a) **2-frame integration:** When the capturing frame rate is *higher* than or equal to the playback frame rate (frame duration: $T_c \leq T_p$), a maximum of two source frames are blended together into a captured frame.



(b) **3-frame integration**: When the capturing frame rate is *lower* than the playback frame rate (frame duration: $T_c > T_p$), at least two frames are integrated into one copy frame.

Figure 2: Frame integration during analog capture

and constant over time. Analog processing, however, can include geometric distortions consisting of affine as well as projective transformations, for example in case that a movie is projected on a screen and captured using a digital movie camera.

For *temporal misalignments*, the key question is whether temporal frame boundaries can be recognized or not. An analog channel is often not capable of passing synchronization information, for example, a screen that is being captured. Here, one captured frame might consist of a linear combination of multiple source frames, depending on the differences between the playback and capture frame rate. When the capturing frame rate is *higher* than or equal to the playback frame rate, a maximum of two source frames are blended together into a captured frame, while a *lower* capturing frame rate might result in a single copy frame being a blending of two or more source frames. Figure 2 visualizes the blending of several source frames into a destination frame. In digital processing, however, we can assume that frames do not generally mix during processing, and in case of changing the temporal resolution (e.g., by encoding with a different frame rate), compressors will most likely drop or duplicate entire frames.

2.3 Benefits of Video Registration

While there is a broad range of watermarking systems that apply information with vastly different methods in numerous embedding domains [28], the following observations are based on principles of communication channels that are used for all watermarking schemes in general. There are significant benefits of video registration that apply to different watermarking aspects:

The most relevant advantage is the *improvement of detection strength* of the watermark for non-blind techniques. A challenge with any robust watermarking technology is to hide the information in a invisible and weak fashion. The underlying video signal must be used as the given information carrier. During extraction, the video signal is noise to the embedded watermarking signal and decreases the signal-to-noise ratio during extraction. If the video signal is eliminated, the remaining signal is the watermark information. While the elimination will not be perfect after the content has been transformed, our system aims for the best possible approximation and contributes as much as possible to the reduction of the video signal.

Figure 3 compares the quality of a watermark extraction with and without video registration. In the sample images, we assume the mark to be embedded in the spatial domain by (visibly) adjusting pixel brightnesses. However, the same results can be observed when using a scheme that applies invisible adjustments (for example, a low-frequency luminance modification). For extraction, we assume the host image content to be noise, and filter it out by calculating the difference between the marked and unmarked image so that the embedded information remains. The top row in Figure 3 (a) visualizes the unattacked watermark extraction. In (b), the content is attacked by geometric distortions and the insertion of noise, and our simple extraction scheme fails. Two effects of registration and alignment can be seen in (c): First, a non-blind watermarking scheme that relies on spatial embedding requires spatially aligned original data to be able to extract the mark. Secondly, when assuming the mark to be extractable in a non-blind fashion (e.g., with human extractors), alignment allows to reduce the SNR ratio of the watermarking signal, leading to an improved extraction. In most cases, the quality of the extracted mark is much lower (d) if the watermarked copy is re-transformed by the inverse transformation and compared to the original image.

Another advantage of registration is the ability to estimate the *reliability of the embedded mark* when analyzing the registered original. The presence of the original allows for recreation of the embedding locations and subsequent analysis of how strong the modifications to the content have been. Weaker embedding locations may have been created by the perceptual model that restricts embedding in certain locations. The knowledge of the original strength of the mark can be estimated much more accurately from the original reference than from the degraded content. This information helps to estimate the detection reliability, it can be used to effectively increase detection performance: redundancy considerations and error correction codes in the embedded information can take confidence values into account, derived from the knowledge of the embedding strength. This information can additionally be used to determine the false positive probability of the result, which is crucial for forensic investigation. Another way to improve detection performance can be accomplished by *using location information* that is available after registration. The identification of the location of the embedded information during extraction is a general challenge for watermarking systems. Some systems provide implicit synchronization; others rely on an additional mark that establishes the location of the actual payload. In either case, if the synchronization is provided to the watermark detector and the position of the payload can be derived, the synchronization information can be omitted, allowing a reduced amount of modifications during embedding. In case of a blind scheme, the synchronization information can be eliminated as a potential weak point that restricts watermark detection if the original is available for comparison.

Figure 3: Comparison of different registration techniques for watermark extraction: unattacked watermark extraction (a), extraction without (b) and with (c) registration, extraction with inverse transformation of the watermarked copy (d). The same effects can be observed on an invisible mark, too.

The challenge of watermark synchronization is exploited by attacks that de-synchronize the content by embedding slight geometric variations (e.g., Stirmark [24]), temporal jitter [23] or flicker in video content, and often watermarking systems are vulnerable to a fairly easy, targeted desynchronization attack. Sophisticated registration like the one presented may reduce the effects of those attacks, making the watermarking information accessible again.

Registration does not only improve security for those specialized attacks, but also for involuntary attacks. Common processing such as cropping, cutting, content merging, resizing, overlaying information (broadcast station logos or subtitles) all potentially cover, remove or misplace parts of the embedded watermark information. Registration provides information on the location of missing embedded information that again can be used to improve detection interpretation and reliability.

The improvement in detection performance, even if marginal, results in significant advantages for the overall performance of the watermarking technology when considering the Oracle attack [27] during which an attacker who is in possession of the detector degrades the content in small iterations until a version is found where the mark is just not readable anymore, and the quality is still at the best possible level. If the availability of the original lowers this threshold, the mark becomes readable again.

3. SEMI-AUTOMATIC SPATIAL AND TEMPORAL VIDEO REGISTRATION

While especially spatial registration is quite well researched, the combination of both spatial and temporal misalignments poses a particular challenge: there is a *two-way dependency* that has to be considered in order to achieve the desired level of precision: A precise spatial matching can only be reached when operating on a pair of frames exactly corresponding to each other in time. On the other hand, a temporal registration that aims at finding frame-exact correspondences

cannot reliably compare frames with high precision unless they are spatially aligned.

This was the motivation develop our toolkit that implements a *semi-automatic* approach to break this dependency and to enable an accurate registration for both spatial and temporal misalignments. The manual assistance is kept to a minimum; it allows a reliable initial synchronization that is the basis for an automatic spatial and temporal registration. The basic idea is as follows: For human perception that can easily understand the semantic content of a video frame, it is a simple task to compare and connect a small number of corresponding frames, even though they are significantly distorted. Having some initial temporal correspondences found and verified by the operator, spatial registration methods can be performed on these to estimate the geometric distortion and find a transformation that maps corresponding pixels of the original and copy frame onto each other. When the geometric misalignment is the same for the entire movie, the mapping can simply be applied to all frames. Actually, this can be assumed to be the case most often since a continuously changing geometric distortion would seriously decrease the pleasure watching a movie. At every point where the geometric misalignment changes, the process can be repeated, and a spatial registration can be performed again on another pair of corresponding frames.

Now that corresponding pixels are known, a temporal matching for all frames is possible, directly comparing frames of the original and the copy. This again is a semi-automatic process, from a rough manual estimation to a precise automatic refinement, as described later in this section.

3.1 Initialization and Spatial Synchronization

This section describes the first step of the registration process in more detail. It is important to recognize that a reliable, initial temporal correspondence is crucial for an accurate spatial registration: If a wrong correspondence is used as a basis (even by one frame only), slight differences such as the ones resulting from camera or object motion will be interpreted as geometric distortions and spoil the synchronization. As mentioned above, usually only very few temporal correspondences are necessary, so this can be easily done by an operator.

We now assume at least one pair of frames that correspond in time, i.e., show the same content, but the copy being somehow geometrically distorted. As already stated before, the task of spatial registration is to map pixels of the distorted frame to corresponding pixels of the original. The spatial model used in our algorithms is capable of handling both geometric distortions from digital processing as well as from projection and acquisition with a handheld camcorder. While digital processing mainly introduces 2D affine transformations only, the latter introduces perspective distortions, too.

In the following, we assume the screen and focal plane of the camera to be planar and neglect the effect of lens distortions. This guarantees that the perspective distortion can be modeled using a *plane-to-plane mapping*. According to [10], this is also a collineation and therefore a projective transform. Since this is linear in projective space, a three-dimensional projective transform can be written as a multiplication with a non-singular matrix $\mathbf{H} = \{h_{jk}\}$, with h_{jk} being the transform parameters. For an n-dimensional projective space $\mathbb{P}^n$, the transform matrix' size is $(n+1)\times(n+1)$, so in case of $\mathbb{P}^2$ there are nine parameters, and a general projective transformation of a point $(x, y, w)^\top$ is defined as:

$$\begin{pmatrix} x' \\ y' \\ w' \end{pmatrix} = \begin{bmatrix} h_{00} & h_{01} & h_{02} \\ h_{10} & h_{11} & h_{12} \\ h_{20} & h_{21} & h_{22} \end{bmatrix} \begin{pmatrix} x \\ y \\ w \end{pmatrix} \quad (1)$$

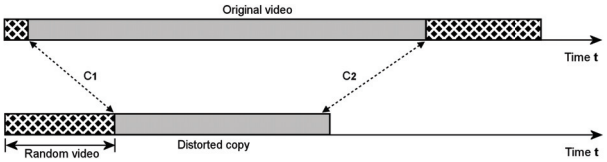
Since homogeneous parameters are scale-invariant, the matrix, too, is only defined up to a scaling factor. Thus, the matrix has only **eight degrees of freedom** in $\mathbb{P}^2$, although it has nine elements. To estimate the transformation between original and copy, the idea is to identify a special set of points $\vec{p_i} = (x_i, y_i, 1)$ in one frame and a corresponding set $\vec{p'_i} = (x'_i, y'_i, 1)$ in the other frame. When inserted into equation 1, each pair of points results in two constraints. An additional constraint can be set in order to remove the scale-invariance. Typically, h_{22} is forced to 1, which normalizes the matrix. As a result, four pairs of corresponding points are sufficient to uniquely determine the transform between the reference and the distorted frame.

These points can either be set manually, or a feature-point detector is used to automatically find characteristic points that can be detected in *both* frames. Currently, the **Shi and Tomasi detector** that comes with Intel's *OpenCV* [2] library is used as the feature point detector. Matching the points found in the reference and distorted frame can be once again done manually, or automatically utilizing *OpenCV*'s pyramidal implementation of the **Lucas Kanade Feature Tracker** [20]. Especially when using a feature point detector and automatic matching, mismatched *outliners* can spoil the results of the transform matrix' computation. Also, a very large number of point-correspondences are found. To select a good subset and dismiss outliers, the RANSAC algorithm (RANdom SAmple Consensus) [12] is applied, before estimating the transform parameters $\{h_{jk}\}$ using *Levenberg-Marquardt* [18, 21]. The matrix H now allows the transformation of each frame into the other.

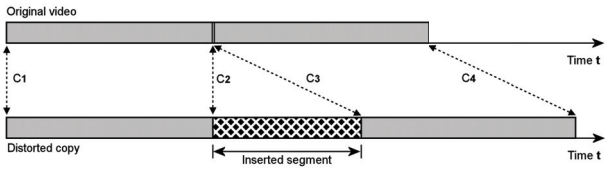
3.2 Temporal Synchronization

Having reached a spatial synchronization, the corresponding pixels are known, and thus the temporal registration can rely on algorithms that directly compare frame contents of the original and the copy instead of comparing features only. On one hand, this means much more computational effort, especially for HD videos; on the other hand, the highest possible precision and reliability can be achieved in this way.

The temporal registration itself is again a semi-automatic two-step process: The first step is a coarse manual synchronization, the second a computational, precise refinement by comparing frame contents within a certain range. For the manual step, the operator links pairs of corresponding frames. Usually, connecting the beginnings and the ends of the videos is sufficient if the copy is not missing any scenes in between. Typical temporal misalignments can easily be described as visualized in Figure 4(a): The diagonally patterned area in the original video is missing in the copy, while the copy contains some random video content in the beginning that needs to be skipped for registration. Also, the content of the distorted copy is drawn with a much shorter bar, visualizing that it consists of fewer frames. Linking the beginnings and ends does not only define the offset of both sequences by *shifting* both start positions, but also defines *temporal scaling*.



(a) Two manually created link keys C_1 and C_2 are sufficient for a coarse temporal synchronization in most cases.



(b) By adding two additional link keys, inserted or missing parts within a video can be modeled.

Figure 4: Rough manual synchronization in time.

When connecting more pairs of frames, inserted or missing parts within a video can be modeled, too (Figure 4(b)). Besides the necessity of the coarse alignment for the following step, it also avoids the need for cutting the movie beforehand (which might further degrade video quality due to decoding/re-encoding).

Assuming the frame rate to be constant, correspondences of all frames can be interpolated from the manually set ones. As the beginnings and ends are reliably linked by the operator, this approximation is already quite accurate, it differs only by a certain amount of w frames from the real correspondences. The task of the computational refinement step is now to precisely *compare* the frames of the original and the copy *within this window of size* w, to find exact correspondences and determine which frames were dropped/duplicated or got blended together. For this, our application supports two models of the temporal distortion, following the misalignments that were categorized into resulting from digital or analog/camcorded reproduction.

3.2.1 Temporal model for digital reproduction

The digital model assumes that *temporal frame boundaries are kept* during all kinds of processing, and thus temporal artifacts are **frame drop or duplication** only. In this case, the similarities between frames within the window around the interpolated match are calculated, and the pair with the highest similarity is considered the corresponding one. Depending on whether the resulting correspondence mapping should go from original to copy or vice versa, one candidate frame from either video is compared to all frames within the window on the other video (see Figure 5). The registration thus tries to minimize the matching error by selecting those frames as corresponding that show the highest similarity. As our similarity measure, we use the **normalized correlation coefficient**:

$$N_{\text{ccoeff}} = \frac{\sum_{x,y}(\tilde{I}(x,y)\tilde{J}(x,y))^2}{\sqrt{\sum_{x,y}\tilde{I}(x,y)^2 \sum_{x,y}\tilde{J}(x,y)^2}} \qquad (2)$$

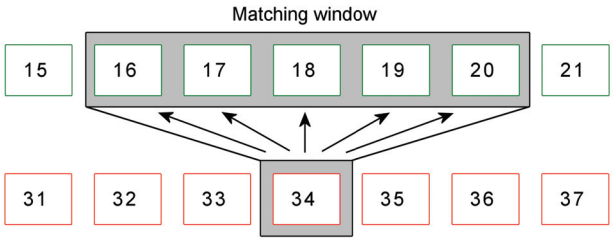


Figure 5: Matching window of five frames.

where $\tilde{I}(x,y) = I(x,y) - \bar{I}$ and $\tilde{J}(x,y) = J(x,y) - \bar{J}$, $\bar{I}$, $\bar{J}$ representing the average pixel value in each image:

$$\bar{I} = \frac{1}{wh}\sum_{x',y'} I(x',y'), \qquad \bar{J} = \frac{1}{wh}\sum_{x',y'} J(x',y'). \qquad (3)$$

Alternative measures are selectable through the GUI, but compared to the often used mean squared error (MSE) the normalized correlation coefficient is robust against changes in brightness and contrast. To some extend, this substitutes a histogram synchronization and is an advantage especially when handling camcorder-recorded videos since the automatic gain control (AGC) and white balance are continously modifying the video's brighness and contrast.

The more useful direction of determining correspondences is **from** the distorted copy **to** the original. This has two reasons:

1. Looking for corresponding frames that are definitely present and not duplicated is much more error-prone than deciding on a corresponding frame *or* its absence. For each frame of the copy, it is guaranteed that exactly one corresponding frame exists in the original.

2. Since only the copy contains watermarking information, the processing will most often go through all copy frames and needs to know the corresponding originals.

However, to increase the reliability we propose to determine correspondences from *both* copy-to-original as well as from original-to-copy frames, resulting in a *bidirectional* mapping. Whenever a pair of frames does not have the same partner of highest similarity in either direction (*although the similarity measure is symmetric*), this is an additional indication for frame drop or duplication. Figure 6 shows an example in which frame B has been dropped while frame D is duplicated. This can be detected by looking at the highest frame similarities, as visualized by the arrows. Whenever two frames do not have the same partner, they cannot correspond, as can be seen for a frame drop at the position marked with *a)*, and for duplication at position *b)*.

On highly degraded content, comparing frames is problematic if frames of the original video are very similar to their neighbors (e.g., on short segments that are completely black). To handle this, the comparison of frames as described above is skipped if the similarity of an original frame to its neighbors is above a certain threshold. Instead, the interpolated correspondence is used, adjusted by a dynamic offset to compensate the average interpolation error. The offset is calculated as the moving average of the previous differences between the interpolated correspondence and the finally selected match.

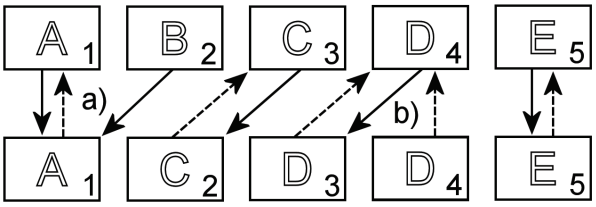


Figure 6: Bidirectional detection of frame drop/repeat. An original video sequence is shown on top with a temporally distorted copy below: Frame B has been dropped while frame D is duplicated.

3.2.2 Temporal model for analog reproduction

The second temporal model supported is a *two-frame integration* model, that is appropriate for misalignments from an analog/camcorded reproduction. Instead of frame drop or duplication only, it assumes copy frames to be a **linear blending of two source frames**, as resulting when temporal frame boundaries are not kept. Again, our algorithm operates on the frames inside a certain window of size w around the interpolated frame correspondences. The result provides the two source frames a copy frames consists of, and also the blending ratio between them. To accomplish this, frame similarities are not calculated between entire frames of original and copy, but between different *blendings* of neighboring source frames and a copy frame. For each comparison the blending ratio with the highest similarity is stored. Obviously, the similarity of each blending to the copy frame will hardly differ if successive frames of the original are already very similar, and the comparison will not yield any meaningful result. Thus, this operation is only performed where the original video has hard transitions such as cuts, or parts with significant motion. Since frames are always assumed to be blended rather than just dropped or repeated, the blending ratios can be interpolated for all frames in between, based on the frame rates of the original and the copy.

3.3 Integration into Watermarking Systems

In order to integrate video registration into a watermarking tool chain, the results need to be passed to a subsequent watermark detector. We support several options to do that:

Export: The easiest way is exporting the results of the spatial and temporal registration to a file, e.g., a CSV text file. All subsequent processing can simply import this file and read the required information (transform matrices, temporal correspondences).

Aligned Original: The second option is to create an output video (or a large number of frames) that is an aligned version of the unmarked original. All spatial and temporal misalignments found in the distorted copy are also applied to the original, so that both original and copy frames correspond to each other. For each frame of the copy video, the detector can now simply read the same frame from the aligned original to have both available side-by-side.

Plug-in: To avoid the detour of creating files, the registration toolkit provides a plug-in system to load dynamically linked libraries. Once a plug-in is installed, it can receive a callback for each processed frame, having access to the copy and the aligned original frame. Since a plug-in can also be integrated at basically any step during the registration process, this method offers the highest level of control.

3.4 User Interface

As the registration presented here is a semi-automatic approach, our implementation provides a rich graphical user interface to assist with all manual tasks. It is intended to be used by professionals, but simplifies the necessary steps as much as possible. In order to directly work on the video's content without delay, it provides a framework to randomly access, display, and process video frames from specified input videos, including the decoding of a wide variety of common video formats and codecs through the Microsoft DirectShow API.

Figure 7 depicts the main application window and gives a first impression of the graphical user interface. At the bottom, a scrollable timeline allows easy navigation through both input videos with random access to all frames. The window's central area can display several different workspaces, each serving a special purpose like video playback, frame display and analysis, or editing feature points. In this screenshot, the most important "Edit" workspace is selected, showing the currently selected frames side-by-side. A selected frame is dynamically loaded in full resolution by simply clicking on its thumbnail. The new Office 2007 Ribbon menu on top holds all necessary controls to work with the application. They are categorized in two tabs, both can be seen in the lower part of Figure 7. The round application button provides commands to save or load a project, export results, or setup general program options. On the right hand side of the main window, different pane windows can be selected through a tab control.

The scrollable timeline provides several features to allow a smooth and convenient operation: First, is allows instant previewing and scrolling through all frames, even for HD videos. When scrolling, thumbnails are dynamically extracted for all currently visible frames in a background thread and presented as soon as they are loaded, not blocking the user interface. Second, there is only *one* scrollbar to both input videos (original and copy) for easy scrolling and navigation. As soon as the user has manually connected both the beginnings and ends, the videos can be automatically aligned so that the visible thumbnails are matching the interpolated correspondences. Thus, when scrolling through the thumbnails, the frames of original and copy will be immediately displayed, roughly aligned in time, independent of the total number of frames.

As described in the previous section, both the spatial and the temporal manual synchronization require the operator to connect a limited number of corresponding frames. A reliable initial correspondence is the basis for the spatial registration, and the temporal registration needs (at least two) manually connected correspondences as a rough synchronization in first place. For the application, both is unified into so-called *link-keys*, linking a pair of frames of original and copy. All link-keys are used to interpolate the temporal correspondences, and the spatial registration can be performed on frames connected by a link-key (and *only* on those). To do so, the "Edit" workspace allows identifying and matching feature points using the algorithms presented.

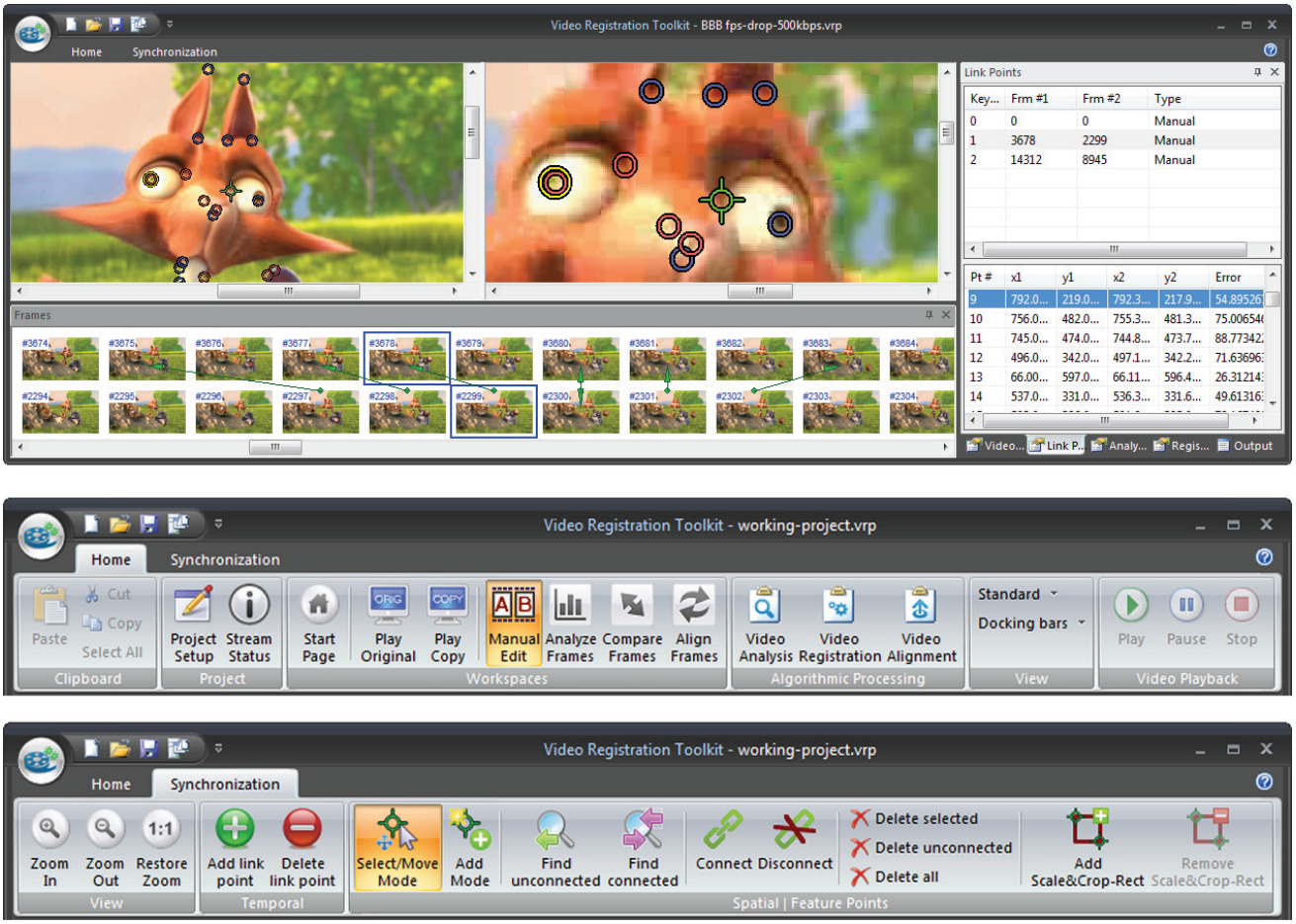


Figure 7: Application screenshot.

In addition, users can also manually add, delete, or modify feature points and correspondences. For precise operation, the view is capable of zooming and panning the frames and feature points displayed. The overall *manual* effort of the operator is indeed relatively small. A typical registration requires the following steps:

1. Open the original and the copy as input videos.
2. Create one or more link-keys and start the automatic spatial synchronization by using feature point detection and matching algorithms.
3. Create two more link keys to connect the beginnings and ends (rough temporal synchronization); more keys are only needed when parts of the video are inserted or missing.
4. Configure and start the automatic temporal registration through a setup dialog window (precise temporal registration).
5. Export or store the results.

Connecting the beginnings and ends can be done first, in order to use the alignment feature of the scrollable timeline. Also, the resulting correspondences are visualized on the timeline by green arrows between the thumbnail frames. Different workspaces also provide the functionality to analyze, geometrically transform and compare selected frames.

4. EXPERIMENTAL RESULTS

Although our synchronization scheme is straightforward, it provides very good results for two reasons:

- We integrate human capabilities where reliable machine-based approaches are hard to employ or might easily fail, although the manual effort is relatively small.
- By direct comparison of frame pixels rather than extracted features, we trade computational cost for a maximum registration accuracy.

Our application scenario is to support the *forensic* extraction of robust watermarks: The readout process is only performed *on-demand* for a limited number of video files and does not require to be operated in an automated environment. Instead, the focus is on a high quality of the registration. Thus, both the human interaction as well as higher computational costs are acceptable measures to achive best results.

Figure 8: An original frame (a) after perspective distortion and compression with a very low bit rate (b). As in camcorded copies, two successive source frames are blended into a copy frame in (c) (with a ratio of 50%).

Task	Processing times
All manual tasks	2-10 minutes
Spatial registration	1-2 seconds
Alignment & temporal registration	at least 5 hours

Table 1: Processing times.

4.1 Performance and Operation

Our software tool was developed for professional users to enable a reliable temporal and spatial video registration. Experiments have shown that all necessary steps can be carried out on a consumer notebook (*Acer Aspire 6930*, with an *Intel Core 2 Duo*, 2GHz) with videos up to full HD resolution (1920x1080). The software is currently used at Verimatrix Inc. as part of their forensic watermarking tool chain and not intended to be released to the public domain, so no large evaluation on the ease of use was conducted at this point.

Table 1 shows typical processing times on a consumer notebook. The computational cost for the temporal registration increases linearly in the number of frames n, but quadraticly when increasing the window size w. However, the interpolated results are usually very good already, so typically not more than one ($w = 3$) or two ($w = 5$) neighboring frames around the interpolation need to be compared. Although the processing is quite expensive, it is mostly image comparisons that can be easily parallelized and/or performed on accelerated graphics hardware.

4.2 Registration Quality

We have conducted a set of experiments to test the correctness and accuracy of the temporal and spatial registration. The computer animated short movie *Big Buck Bunny*[1] (854x480p, MPEG4, 24 fps, 2500 kbps, 9:56 min, 14,315 frames) and a digitalized VHS video from the University of Mannheim (672x560, MPEG4, 25 fps, 798 kbps, 5:41 min, 8,547 frames) have been used as reference videos. In addition, we have extracted a short clip of 80 frames from the Big Buck Bunny movie, scaled to 640x360, and used it for a detailed analysis. The temporal registration is especially challenging for the short clip because all frames are very similar (average normalized correlation coefficient is $0,9389$ with a standard deviation of $0,0461$).

For the tests, spatial and temporal distortions have been applied to the test videos, and distorted versions of the reference clips have been created. In case of the short clip, all combinations of distortions were analyzed: perspective (see Figure 8), cropping, random frame drop (5-10%), low bit rate compression (170-854 kbps), and random frame blending as caused by analog capturing. Table 2 lists all test clips and their properties.

[1](c) copyright Blender Foundation, www.bigbuckbunny.org

Clip	Distortion	Frames	kbps
REF1	*Big Buck Bunny*	80	1027
	all combinations	72–80	170–854
REF2	*Big Buck Bunny*	14,314	2500
	resized to 640x480	14,314	2123
	reduced to 15fps	8,947	1789
	Camcorded (persp., frame blending, ...)	15,061	5010
REF3	*Univ. of Mannheim*	8,547	798
	random frame drop	7,966	798

Table 2: Test videos for the evaluation.

First, the **false positive rate** of the temporal registration was analyzed, to see if comparing a reference frame against all frames within the registration window always favors the correct correspondence. To conduct this experiment, no temporal modification is applied to the copy, so frames of original and copy exactly correspond already. However, spatial distortions might be applied. When original and copy are now registered to each other, any frame drop or duplication that is found is a *false positive* since the videos are temporally identical. For our tests, the size of the window for comparison was set to five frames around the interpolation, which was intentionally biased by one frame. Another parameter is the threshold s that decides when the registration is skipped: This happens if the correlation of the neighboring frames is above this threshold. For the short clips that are only spatially distorted, we used a very high threshold of $s = 0.9999$, so the comparison was never skipped. Still, no frame drops or duplications where detected by mistake. The same test was conducted on the entire Big Buck Bunny movie. When registered with itself (no attack at all), for five of the 14,314 frames a frame drop was wrongly detected when skipping the registration with a threshold of 0.999. When resizing the copy to 640x480 with change of the aspect ratio, the spatial alignment introduces a source of imprecision: the number of false positives increases to eight. When the threshold parameter is additionally set to 0.9999, also frames that are almost identical are compared (e.g., on fades to black), resulting in twelve drops detected by mistake.

As a second test, the **frame drop detection** was performed on the short clip where different distortions (*perspective transform, cropping, low bit rate compression*) and

all combinations of these could be evaluated. For all tested clips, every single drop was successfully detected. To analyze the frame drop detection on a longer video, 581 frames have been randomly dropped from the University of Mannheim video, their positions stored, and finally compared to the results from the registration. 0,0119% of all dropped frames were not detected, but most of them being in areas with almost identical neighboring frames.

To verify the algorithm for **frame blending**, several frames were synthetically blended as caused by analog capturing. All locations were found and the correct ratios detected, while no blending was reported for the other frames. Evaluating the quality of the blending detection on real-world data such as a camcorded video is much harder, since no information is reported on the actual blending ratios when the copy is recorded. However, the Big Buck Bunny movie was captured with a consumer camcorder (*Sony DCR-TRV60*), and correct results could be proofed in spot tests.

To test the quality of the **spatial registration**, the normalized *mean squared error* was accumulated over all corresponding frames of the entire video, once with and once without the spatial registration. For the perspectively distorted short clip, the summed up normalized error is 0.2909 without spatial registration, and 0.0059 when an automatic spatial alignment is performed. If low bit rate compression is additionally applied, this cannot be compensated for: The accumulated error with registration increases to 0.0327 then.

5. RELATED WORK

Registration of images and videos is a fundamental task in image processing, and central components of applications like panoramic images [4, 6, 22, 14], video retargeting [16], optical inspection [13], multi-camera capturing [26], video object segmentation [11] or medical imaging [5, 25] require reliable image or video registration techniques. Although most publications focus on spatial registration techniques (e.g., [3, 1, 17]), a large number of temporal registration techniques have also been published [8, 9, 15]. In most cases, the authors assume that only one kind of misalignment (spatial *or* temporal) has to be corrected. While each of these registration operations are useful, the combination of both misalignments poses a particular challenge, due to the interdependence of the spatial and the temporal registration.

Another possible approach than the one presented here is a temporal registration based on a feature profile only. *Delannay* [9] proposed a temporal alignment by matching several key frames only, and interpolating correspondences in between. A key frame is given when the luminance histogram of a frame deviates sufficiently from its predecessor. However, this approach fails when too many key frames are suppressed, or in segments with high motion activity: Here, the key frame selection is unlikely to select exactly the same frames for original and copy.

A similar approach is presented by *Chupeau* [8]: Videos are reduced to a continuous one-dimensional "temporal profile", its values resulting from the distances between the color histograms of successive frames. In a second step, frames are matched based on this profile using dynamic programming.

Unfortunately, neither of these approaches presents a complementing spatial registration. An integrated approach for spatial, temporal and histogram registration (STH) can be found by *Cheng* and *Isnardi* in [7], which is used in the forensic watermarking system offered by the Sarnoff Corporation. In their solution, misalignment in each domain is modeled separately first, and then put together in a combined equation. Afterwards, they try to find parameters for spatial, temporal and histogram alignment to this model that minimize the accumulated mean squared error between original and copy for all frames. Since there is no closed-form solution (the spatial-temporal dependency also exists in a combined model), two sets of parameters (e.g., for temporal and histogram alignment) have to be *fixed* in order to allow optimization of the third one. An iterative approach is chosen where parameters are fixed in turns. Although "fixing" parameters in the first iteration practically ignores distortions in that domain, the idea is that a correct initial guess can still be achieved. However, this is not guaranteed when spatial distortions are not constant over time.

Unfortunately, we could not directly compare this approach with our solution, as it is (like ours) not publicly available. An advantage is the integrated histogram registration, that compensates any color modifications. In our approach, this is partly achieved by using the normalized correlation coefficient as the similarity measure, which is robust against changes in brightness and contrast. Also, an integrated iterative optimization might run into a local optimum, while in our solution spatial and temporal parameters are optimized separately and only *one* iteration is necessary, based on the manual interpolation. (This also limits the computational cost.) Another drawback of a model unifying all misalignment is the fact that fewer helpful constraints can be derived, especially regarding the temporal registration: We have defined two models: (1) digital processing which causes frame drop or repeat as major temporal misalignment, and (2) a model for videos which typically include an analog processing step (e.g., projecting the video onto a cinema screen) where two or more frames are blended together during the capture.

6. CONCLUSIONS AND OUTLOOK

We have developed a software tool for the semi-automatic registration of videos. It allows the synchronization of a misaligned copy of a video with its original for each frame and pixel. Our implementation breaks the interdependence between temporal and spatial alignment with subsequent optimization and integration of human abilities in order to improve detection of forensic digital watermarks. We employ an efficient scheme to allow user contribution to machine processing: Human interaction is crucial for reliability and performance while machine processing helps to complete and use the results in an effective way.

During development, an emphasis has been placed on a fast and responsive graphical user interface for a convenient manual synchronization on the one hand, and an efficient multi-threaded implementation of the processing on the other hand, utilizing the power of modern multi-core processors. Thumbnails of all input video frames are displayed on a timeline and are loaded dynamically when scrolling, allowing to edit temporal as well as spatial correspondences for selected frames on feature-length HD content.

Possible further improvements to the presented spatial registration result can be expected from the use of transform-invariant feature point algorithms (SIFT [19]). Image transformations and the temporal registration through image comparisons are perfect applications for hardware acceleration through general purpose GPU programming (GPGPU).

Another area where we see a potential improvement is the integration of digital fingerprinting approaches to identify corresponding sections of the original and copy. While we believe that this will aid in reducing the required human interaction in many cases, we also believe that it will not eliminate the need for human assistance when aiming for maximum registration performance. When considering current approaches, the human capabilities are superior for this application in particular when considering targeted attacks by a human adversary.

7. REFERENCES

[1] S. Baudry, P. Nguyen, and H. Maitre. Estimation of geometric distortions in digital watermarking. In *ICIP, Rochester, NY, USA*, pages II–885 – II–888, 2002.

[2] G. Bradski and A. Kaehler. *Learning OpenCV*. Learning OpenCV, 2008.

[3] L. G. Brown. A survey of image registration techniques. In *ACM Computing Surveys*, volume 24(4), pages 325–376. ACM Press, Dezember 1992.

[4] M. Brown and D. G. Lowe. Automatic panoramic image stitching using invariant features. In *International Journal of Computer Vision*, volume 74 (1), pages 59–73. Kluwer Academic Publishers, August 2007.

[5] X. Cao and Q. Ruan. A survey on evaluation methods for medical image registration. In *IEEE/ICME International Conference on Complex Medical Engineering*, pages 718–721, May 2007.

[6] H. Chen. Gradient-based approach for fine registration of panorama images. In *Journal of Computer Science and Technology*, volume 19 (5), pages 691–697. Springer, September 2004.

[7] H. Cheng and M. A. Isnardi. Spatial temporal and histogram video registration for digital watermark detection. In *Proc. of International Conference on Image Processing*, volume 2, pages II – 735–8, September 2003.

[8] B. Chupeau, L. Oisel, and P. Jouet. Temporal Video Registration for Watermark Detection. In *ICASSP, Toulouse, France*, volume 2, 2006.

[9] D. Delannay, C. de Roover, and B. M. M. Macq. Temporal alignment of video sequences for watermarking systems. In *Proc. of SPIE conference on Security and Watermarking of Multimedia Contents V*, volume 5020, pages 481–492, 2003.

[10] Dirk Sven Farin. *Automatic Video Segmentation Employing Object/Camera Modeling Techniques: Dissertation*. PhD thesis, Technische Universiteit Eindhoven, Eindhoven, 2005.

[11] D. Farin, T. Haenselmann, S. Kopf, G. Kühne, and W. Effelsberg. Segmentation and classification of moving video objects. In B. Furht and O. Marques, editors, *Handbook of Video Databases: Design and Applications*, volume 8 of *Internet and Communications Series*, pages 561–591. CRC Press, Boca Raton, FL, USA, September 2003.

[12] M. A. Fischler and R. C. Bolles. *Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1987.

[13] B. Guthier, S. Kopf, and W. Effelsberg. High-resolution inline video-aoi for printed circuit assemblies. In *Proc. of IS&T/SPIE conference on Image Processing: Machine Vision Applications II*, volume 7251, January 2009.

[14] T. Haenselmann, M. Busse, S. Kopf, T. King, and W. Effelsberg. Multi perspective panoramic imaging. In *Image and Vision Computing*, volume 27 (4), pages 391–401, March 2009.

[15] Hui Cheng. Temporal registration of video sequences. In *Proc. of International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages 489–492, April 2003.

[16] S. Kopf, J. Kiess, H. Lemelson, and W. Effelsberg. FSCAV: Fast seam carving for size adaptation of videos. In *Proc. of ACM International Conference on Multimedia*, pages 321–330, 2009.

[17] R. Kumar, H. Sawhney, J. Asmuth, A. Pope, and S. Hsu. Registration of video to geo-referenced imagery. In *Proc. of International Conference on Pattern Recognition*, volume 2, pages 1393–1400, Aug. 1998.

[18] K. Levenberg. A method for the solution of certain problems in least squares. *Quart. Applied Math.*, 2:164–168, 1944.

[19] D. G. Lowe. Distinctive image features from scale-invariant keypoints. In *International Journal of Computer Vision*, volume 60(2), pages 91–110. Kluwer Academic Publishers, November 2004.

[20] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision (DARPA). In *Proc. of the 1981 DARPA Image Understanding Workshop*, pages 121–130, April 1981.

[21] D. W. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*, 11(2):431–441, 1963.

[22] A. Mills and G. Dudek. Image stitching with dynamic elements. In *Image and Vision Computing*, volume 27 (10), pages 1593–1602, September 2009.

[23] F. Petitcolas, R. Anderson, and M. Kuhn. Attacks on copyright marking systems. In *Second workshop on information hiding*, volume 1525, pages 218–238, April 1998.

[24] F. A. P. Petitcolas. Watermarking schemes evaluation. In *IEEE Signal Processing*, volume 17(5), pages 58–64, September 2000.

[25] J. Pluim, J. Maintz, and M. Viergever. Mutual-information-based registration of medical images: A survey. In *IEEE Transactions on Medical Imaging*, volume 22 (8), pages 986–1004, Aug. 2003.

[26] S. Vedula, S. Baker, and T. Kanade. Image-based spatio-temporal modeling and view interpolation of dynamic events. In *ACM Transactions on Graphics*, volume 24 (2), pages 240–261, April 2005.

[27] I. Venturini. Counteracting oracle attacks. In *Proceedings of the 2004 workshop on Multimedia and security*, pages 187–192, 2004.

[28] D. Zheng, Y. Liu, J. Zhao, and A. E. Saddik. A survey of RST invariant image watermarking algorithms. In *ACM Computing Surveys*, volume 39 (2), 2007.