THE USE OF FACE INDICATOR FUNCTIONS FOR VIDEO INDEXING AND FINGERPRINTING

Costas Cotsaces, Nikos Nikolaidis and Ioannis Pitas

Department of Informatics
Aristotle University of Thessaloniki
Thessaloniki 54124, Greece
(email: pitas@zeus.csd.auth.gr).

ABSTRACT

The characterization of a video segment with a digital signature is a fundamental task in video processing. It is necessary for video indexing, copyright protection and other tasks. Semantic video signatures are those that are based on high-level content information rather than on low-level features of the video stream. The major advantage of such signatures is that they are invariant to nearly all types of distortion. A major semantic feature of a video is the appearance of specific people in specific frames. Because of the great amount of research that has been performed on the subject of face detection and recognition, the extraction of such information is generally tractable. We have developed an indexing and retrieval method that uses the pre-extracted output of face detection and recognition to perform fast semantic indexing and retrieval of video segments. The biggest advantage of our approach is that the evaluation of similarity is convolution-based, and is thus resistant to perturbations in the signature and independent of the exact boundaries of the query segment.

1. INTRODUCTION

Video indexing [2] requires that a greatly reduced dimensionality representation (a *signature*) is computed for each video segment. We call this representation a *signature*. The task of video fingerprinting is very similar to video indexing, as it also involves searching for videos based on some signature, its main difference being that the comparison of the signatures of two videos needs to be stricter.

In this work, we take advantage of a specific type of semantic information, namely information about the appearance of faces of distinct individuals, in order to characterize a video segment in a robust way. We do not concern ourselves with the extraction of face-related information, since ample work has been performed on the subject. This work tries to solve the problems of consistency and robustness with regards to face-based indexing, to represent face information with minimal redundancy, and also to find a fast (logarithmic-time) search method.

Using face-related information for video indexing and/or fingerprinting is not a new idea. However, all works until now [1, 3, 4, 5, 6] have focused on the extraction of the face-related information and not on its organization and efficient indexing. In effect, they are works on face recognition with a view to application on indexing. As such, they actually present an excellent foundation in providing input to the current work. This is especially true for the works of Satoh [4] and of Eickeler et al [3], who perform identity recognition on the faces they detect. In our work we do not propose a face detection and recognition method, but we investigate the effect of different parameters of the face detection and recognition process on the indexing performance of our method. The data used for this purpose is constructed by a probabilistic model describing the appearance of faces in videos.

The paper is organized as follows: in section 2 we give a theoretical overview of our algorithm, in section 3 we describe the data we use for experimental verification, section 4 provides the experimental results, and conclusions are presented in the last section.

2. ALGORITHM DESCRIPTION

2.1. Format of Signature

Let $\mathbf{V} = \{f_1\ f_2\ \ldots\ f_N\}$ be a video consisting of a number of consecutive frames $f_n, n = 1 \ldots N$. that we wish to characterize through an appropriately constructed signature. Let $\mathbf{S} = \{s_1\ s_2\ \ldots\ s_M\}$ be the set of all the individuals $s_m, m = 1 \ldots M$ that have been imaged in the video. Optionally, with no loss of generality, we can assume $\mathbf{S}$ to contain only the individuals of interest. This can mean, for example, excluding the extras in a motion picture.

This work has been supported by the European Union project MUSCLE: Multimedia Understanding through Semantics, Computation and Learning,

ISBN: 952-15-1364-0

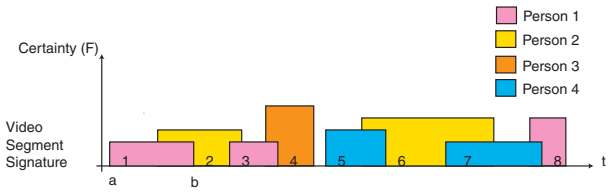

Fig. 1. Example of the characterization of a video segment by quartets. Colors correspond to distinct individuals. Signature elements are represented by numbered rectangles.

Let us then assume a face detector and recognizer F whose output is the certainty:

$$G(n,m) = P\{s_m \text{ is imaged in } f_n\}$$

The face detector/recognizer can either be hard, in which case $G(n,m) \in \{0,1\}$ or fuzzy, in which case $G(n,m) \in [0,1]$.

For each person s_m it is then possible to find all intervals $I_i^m = [a_i^m, b_i^m]$ such that $G(n,m) > 0,\ n \in [a_i^m, b_i^m]$ and $I_i^m \not\subset I_j^m, \forall i \neq j$. Using I_i^m we can then define a *face occurrence* $F_i^m = \overline{F(n,m)}|_{n=a_i^m}^{b_i^m}$ as the average certainty within the interval I_i^m, that a specific person is imaged. So we can approximate $G(n,m)$ with

$$F(n,m) = \sum_i F_i^m \left[u(n-a_i^m) - u(n-b_i^m)\right] \quad (1)$$

where $u(n)$ is the unit step function and $[a_i^m, b_i^m]$ is the i-th interval that contains the face of the m-th person.

Therefore, the video $\mathbf{V}$ is characterized by quartets of values $(s_m, F_i^m, a_i^m, b_i^m)$. Each quartet corresponds to a unique face appearance, i.e. it conveys the information that person s_m has been detected from frame a_i^m to frame b_i^m with a confidence of F_i^m. An example is given in Figure 1. The number of quartets in a video is equal to $\sum_{m=1}^{M} g_m \ll N \times M$, where g_m is the number of appearances of person s_m in the video, and N and M are the total numbers of frames and persons in the video.

2.2. Signature Extraction

The extraction of the signature from the video can be simply a matter of applying the procedure described in the above section to the output of the face detection/recognition module. In practice, in order to reduce the amount of redundant data in the signature, it is better to discard face occurrences that are too short and to unify proximate occurrences of the same face. T

2.3. Signature Similarity

In order to compare two signatures $F_1(n,m)$ and $F_2(n,m)$, which are extracted from two videos $\mathbf{V}_1$ and $\mathbf{V}_2$, and which refer to a common set of faces $\mathbf{S}$. In the case of a hard detector, the optimal displacement d between the two sequences is the one that gives the highest face co-occurrence:

$$P_{\textit{co-occurrence}} = \sum_{n=1}^{N}\sum_{m=1}^{M} \frac{F_1(n,m)\cdot F_2(n+d,m)}{NM}$$

On the other hand in the case of a fuzzy detector the probability of co-occurrence in a frame is the minimum of the two outputs of the detector. This can be explained by the fact that a lower certainty in one video overrides the greater certainty in the other video. Therefore we have:

$$P_{\textit{co-occurrence}} = \sum_{n=1}^{N}\sum_{m=1}^{M} \frac{\min(F_1(n,m), F_2(n+d,m))}{NM}$$

2.4. Search-Matching Algorithm

When the database is initialized, an index I_{sa} is created over all the signature quartets $\mathbf{Q}^{db} = (s^{db}, F^{db}, a^{db}, b^{db})$ in the database, indexing them first on the face s and then on the start frame a. It is assumed that the videos are arranged sequentially in the database. Two other indexes I_a and I_b are created based on start frame a and end frame b alone.

In the following, when we declare a sub- or super-scripted Q, we will assume it is a quartet of the form $Q = \{s, F, a, b\}$, where s, F, a and b have the same sub- and super-scripts as Q.

The following algorithm (illustrated in Figure 2) is proposed for finding matching segments in the database with respect to a query segment $\mathbf{V}_{\textit{query}}$:

1. Choose the n first quartets $\mathbf{Q}^{\textit{first}} = \{Q_1^{\textit{first}} \ \ldots \ Q_n^{\textit{first}}\}$ in the query signature where $n = c_1$ is a predefined constant. This is the part of the signature that is going to be used for searching the database.

2. Find the most trustworthy quartet $Q^{\textit{trust}} \in \mathbf{Q}^{\textit{first}}$, i.e. the one that has the greatest area:

$$F^{\textit{trust}}(b^{\textit{trust}} - a^{\textit{trust}}) = \max_j F_j^{\textit{first}}(b_j^{\textit{first}} - a_j^{\textit{first}})$$

3. Find through the index I_{sa} all quartets $\mathbf{Q}^{\textit{fit}}$ in the database that refer to the same person as $Q^{\textit{trust}}$:

$$\mathbf{Q}^{\textit{fit}} = \{Q^{\textit{fit}} \in \mathbf{Q}^{db} : s^{\textit{fit}} = s^{\textit{trust}}\}$$

Then rank them based on their start frame $a^{\textit{fit}}$.

4. For each $Q_i^{\textit{fit}} \in \mathbf{Q}^{\textit{fit}}$:

   (a) Add the pair of quartets $Q_i^{\textit{trust}}, Q_i^{\textit{fit}}$ into a new list L.

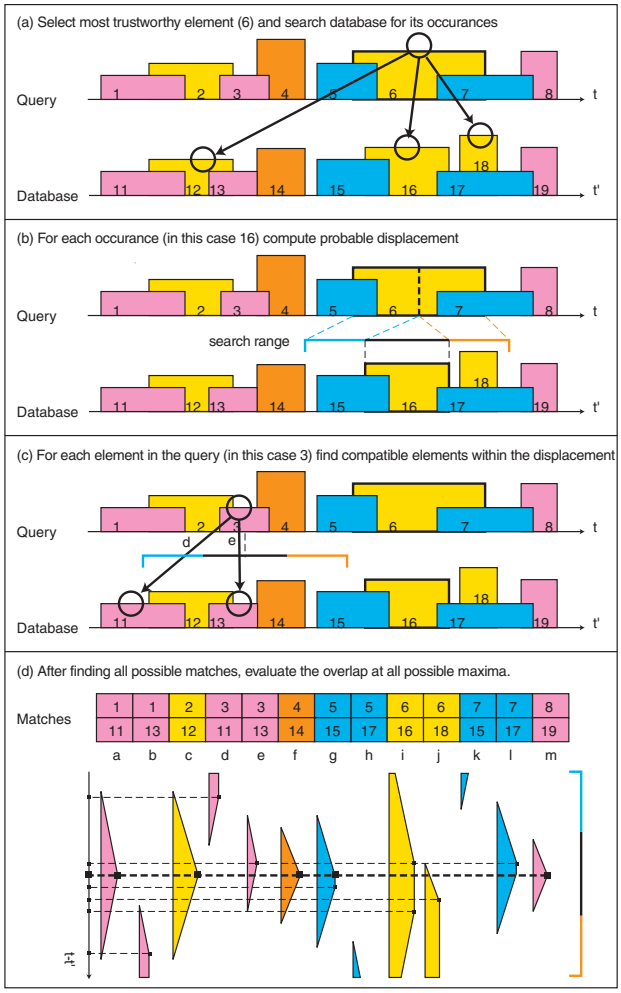

Fig. 2. Graphical overview of the algorithm. Colors correspond to distinct individuals. Signature elements are represented by numbered rectangles.

(b) Calculate the displacement window $[a_i^{disp}, b_i^{disp}]$ within which a possible match may be found in the database , where:

$$b_i^{disp} = \frac{(b^{fit} - a^{fit})}{2} + \frac{(b^{trust} - a^{trust})}{2}$$

$$a_i^{disp} = -b_i^{disp}$$

(c) For each query quartet $Q_j^{first} \in \mathbf{Q}^{first}$:

i. Find, through the database indexes I_a and I_b, the set of compatible quartets $\mathbf{Q}_{ij}^{comp}$ in the database, i.e. quartets that belong to person s_j^{first} and which overlap with a window of size $b_i^{disp} - a_i^{disp}$ which is centered on Q_m^{fit}.

ii. If $\mathbf{Q}_{ij}^{comp}$ is empty, increment a counter n_1. If $n_1 > T_{reject}$ where T_{reject} a threshold, proceed to the next compatible database quartet Q_{i+1}^{fit}.

iii. Add the pairs consisting of Q_j^{first} and all $Q_{ij}^{comp} \in \mathbf{Q}_{ij}^{comp}$ into the list L. It should be noted that since a face cannot exist more than once in each frame, the intervals in $\mathbf{Q}_{ij}^{comp}$ do not overlap.

(d) Evaluate the area of overlap v_{il} of all pairs Q^{query}, Q^{data} in list L for all displacements d_{il} that correspond to possible maxima. These displacements are those that $a^{query} + d_{il} = a^{data}$ or $b^{query} + d_{il} = b^{data}$. Select the maximum match quality $v_i^{optimal} = \max_l v_{il}$ and also keep the corresponding displacement $d_i^{optimal}$.

(e) Clear list L.

5. Select $v^{optimal} = \max_i v_i^{optimal}$. If this is above a threshold T_v (which depends on the size of $\mathbf{Q}^{first}$), then declare a match, otherwise declare no match. Also keep the corresponding displacement $d^{optimal}$.

6. Optionally, if no match is found repeat all above for the next most trustworthy quartet. In our experiments we have done so.

Note that if one wishes to continue verifying the retrieved segment, he can repeat 4c for the quartets beyond $\mathbf{Q}^{first}$, keeping $d^{optimal}$ but adjusting $v^{optimal}$ and checking if it exceeds a modified threshold T_v.

3. DERIVATION OF EXPERIMENTAL DATA

The interest of this work is investigating the performance of the proposed video indexing and fingerprinting method when applied on large databases (typically containing hundreds of hours of video). However, the effort of applying different types of face detectors and recognizers on large video databases, in order to derive the data required for the experimental performance evaluation of the proposed indexing and fingerprinting method, is extremely high. Therefore we have elected to perform the experimental testing of our algorithm on appropriately constructed artificial data. To achieve that we have formulated a probabilistic model which describes the ground truth of the appearance of faces in videos, and a second probabilistic model which describes the behavior of the face detection and recognition module. This approach has the advantage that we can easily test our algorithm on videos and face detection and recognition methods that have different characteristics, by varying the parameters of the model. In the case of the ground truth model,

the model parameters can be obtained by manually annotating a small corpus of video data. In the case of the face recognition and detection model, the model parameters can be obtained by running the respective algorithms on a sufficient set of data, but in our experiments we have varied them to explore the behavior of our method with respect to algorithms with different characteristics.

3.1. Ground Truth Modelling

We model the appearance of persons in a video by considering the fact that the video is inherently composed of scenes, which are in turn composed of shots. Since scenes are spatio-temporally continuous in the context of the depicted world, and shots are spatio-temporally continuous in the video domain, we can assume different probabilities of appearance of a specific person for each scene and shot.

In order to construct the above model we needed three sets of information:

1. *The structure of the model*, i.e. the random variables it contains and their interrelations. This was constructed by analyzing the motion picture production process.
2. *The specific probability distributions of the random variables appearing in this model*. To estimate these, we have first manually annotated a moderately large corpus of video data by marking the faces appearing in it, and also the scene and shot boundaries present. We then tried to find appropriate distributions by using a combination of statistical testing (Kolmogorov – Smirnoff fit tests) and analysis of the physical meaning of the variables.
3. *The parameters of the above distributions* (mean, standard deviation etc). These were again computed from the manually annotated video data.

In the following we give only a brief outline of the model used:

1. Prevalence of each person in the video.
2. Probabilistic model of the scenes and shots of the video. Three random variables are used for this purpose:
   (a) The number of shots in a scene.
   (b) The length of each shot.
   (c) The average distance of persons from the camera in each shot.
3. Number of persons appearing in each scene.
4. Prevalence of each person in the scene
5. Number of persons in each shot.
6. Prevalence of each person in the shot
7. Signature quartets. In order to describe signature quartets four random variables are needed. These are, person identity (s), certainty (F), starting frame (a) and ending frame (b).

For modelling the behavior of a face detector and recognizer, we assume that they introduce inaccuracies according to the random distributions of the following occurrences, which act as noise on the ground truth described above.

1. That the bounds of a signature element will be misdetected. This can be due to periodicity in the initialization of the tracker, or simply partial detection failure.
2. That a face will be misclassified as belonging to a different person. This is the only failure that is attributable to the face recognizer.

4. EXPERIMENTAL RESULTS

Two sets of experiments were performed, one for assessing computational performance, and one assessing robustness with respect to face detection/recognition errors.

4.1. Computational Performance

In order to evaluate the computational performance of our algorithm, we created artificial video databases of different sizes as described in the previous section. Each database consisted of a number of videos, each having a duration of 60 minutes and containing between 1000 and 2000 quartets. The number of different persons for each database was chosen to be 10 times the number of videos. We then selected query segments with an average length of 2.5 minutes and ran our search algorithm on these segments, using a commercial RDBMS system for the implementation. The average times of retrieval are given in table 1.

As it can be seen, the performance of the algorithm is near-logarithmic with respect to the size of the database.

Tab. 1. Average search times

Number of Videos	Search time
100	2 seconds
1000	5 seconds
10000	14 seconds

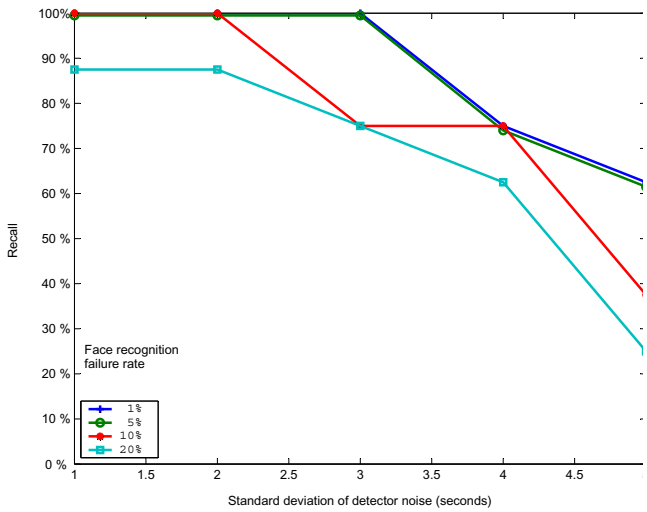

Fig. 3. Recall as a function of noise. The different lines correspond to different failure rates of the recognizer.

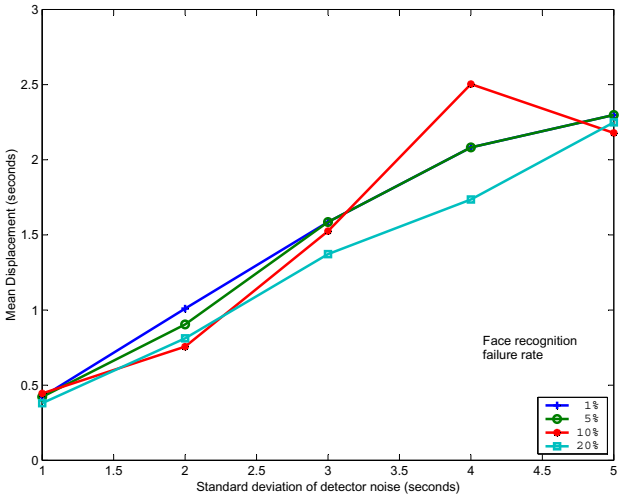

Fig. 4. Accuracy as a function of noise. The different lines correspond to different failure rates of the recognizer.

4.2. Precision and Accuracy

Given that neither face detection algorithms, nor face recognition algorithms are perfect, we performed a series of experiments to test the behavior of our algorithm in the presence of noise introduced during detection and recognition of the faces in the video. The following types of noise have been considered:

1. Change of quartet bounds. This is one of the most typical errors made by face detectors and trackers. Exponential noise has been added to the start time of a quartet (to simulate a delayed detection), and zero mean gaussian noise to the end time of the quartet (to simulate either early loss or false continuation of tracking). It should be noted that when the noise is high it can result in the complete elimination of a quartet. The μ (in the case of a quartet's start frame) and the standard deviation (in the case of a quartet's end frame) of the noise was varied from 1 to 5 seconds.

2. Change of the person's identity in a quartet. This is a typical error made by face recognizers. Here we assumed a chance (between 1% and 20%) that a person's identity would be changed to another random one.

This set of experiments was run using an artificial signature database of 1000 videos, each 60 minutes long. From this database we extracted 8 segments with an average duration of 2.5 minutes, added noise, and we proceeded to search for them in the database. The recall of our algorithm with respect to detector and recognizer noise is shown in figure 3. We can see that for low noise levels the performance is almost perfect. For high noise levels, especially for detector noise, the performance drops dramatically. However we should note that the average quartet length (both in real and artificial videos) is less than 4 seconds, while the added noise (change in the start and end frames) had a standard deviation that was as much as 5 seconds.

Figure 4 shows the average accuracy of the localization of the segments in the database. In other words, this figure depicts the average difference (displacement) between the true position and the estimated position of the query segment. One can see that the displacement is approximately equal to half the noise mean, a fact that proves that the accuracy of the method is very satisfactory.

5. CONCLUSIONS

A method for performing fast indexing, retrieval and fingerprinting of video segments based on the output of face detectors and recognizers has been presented. The proposed method is both robust because it is based on a convolution-like similarity computation, and fast because it makes extensive use of database indexes. Experimental results were computed on artificial data based on realistic models of the appearances of faces in videos and o face detector/recognizer behavior. The results verified that the proposed method performs very satisfactorily, both in terms of computational search efficiency (even in a database of 10000 hours of video), and in terms of recall and accuracy.

6. REFERENCES

[1] Yin Chan, Shang-Hung Lin, Yap-Peng Tan, and S.Y. Kung. Video shot classification using human faces. In *IEEE International Conference on Image Processing (ICIP)*, volume 3, pages 843–846, 1996.

[2] Nevenka Dimitrova, Hong-Jiang Zhang, Behzad Shahraray, Ibrahim Sezan, Thomas Huang, and Avideh Zakhor. Applications of video-content analysis and retrieval. *IEEE Multimedia Magazine*, 9(3):42 – 55, July 2002.

[3] Stefan Eickeler, Frank Wallhoff, Uri Iurgel, and Gerhard Rigoll. Content-based indexing of images and video using face detection and recognition methods. In *IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2001.

[4] Shin'ichi Satoh. Comparative evaluation of face sequence matching for content-based video access. In *Proc. of the 4th Intl Conf. on Automatic Face and Gesture Recognition(FG2000)*, pages 163 – 168, 2000.

[5] M. Viswanathan, H.S.M. Beigi, A. Tritschler, and F. Maali. Information access using speech, speaker and face recognition. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 493–496, July-August 2000.

[6] Gang Wei and Ishwar K. Sethi. Omni-face detection for video/image content description. In *Proceedings of the 2000 ACM workshops on Multimedia*, Nov 2000.