Selecting the Best Faces to Index Presentation Videos *

Michele Merler
Department of Computer Science
Columbia University
New York, NY 10027
mmerler@cs.columbia.edu

John R. Kender
Department of Computer Science
Columbia University
New York, NY 10027
jrk@cs.columbia.edu

ABSTRACT

We propose a system to select the most representative faces in unstructured presentation videos with respect to two criteria: to optimize matching accuracy between pairs of face tracks, and to select humanly preferred face icons for indexing purposes. We first extract face tracks using state-of-the-art face detection and tracking. A small subset of images are then selected per track in order to maximize matching accuracy between tracks. Finally, representative images are extracted for each speaker in order to build a face index of the video. We tested our approach on 3 unstructured presentation videos of approximately 45 minutes each, for a total of a quarter million frames. Compared to the standard min-min approach, our method achieves higher track matching accuracy (94.22%), while using 6% of the running time. Using an optimal combination of 3 user preference measures, we were able to build face indexes containing 54 speakers (out of the 58 present in the videos) indexing into 795 detected tracks.

Categories and Subject Descriptors H.3 [Information Storage and Retrieval]: H.3.1 Content Analysis and Indexing; D.2.8 [Content representation]:
General Terms Algorithms, Human Factors
Keywords Video Indexing, Face Tracking and Matching

1. INTRODUCTION

Various systems have been proposed to automatically index professional videos such as movies, TV shows, and news based on characters appearing in them [4]. Likewise, in the surveillance and biometrics fields, there is great interest in building high quality compact yet complete face logs from large unstructured static camera footage, to be presented to analysts for further processing[10]. Our work lies at the convergence of these two trends, as it aims at indexing unstructured presentation videos based on speaker appearances, and uses quality measures to select representative face images.

Not much attention has been given instead to unstructured videos shot in uncontrolled environments. The determination of the "best" faces for recognition purposes (especially in terms of pose) has been studied both from a psychological [3] and computational point of view [8].The results suggest that humans are able to infer the full 3D information about the head of a person when seeing a 3/4 view of their face. Liu et al. [8] analyzed the same problem from the computer vision point of view and verified that a 32° pose provides the best generalization performance for face matching algorithms. In the surveillance domain, Fourney et al. [5] present a series of quality measures (based on pose,illumination,sharpness, resolution and skin) to reduce face tracks into face logs.

To the best of our knowledge, the only work trying to assess the quality of speakers video indexes for unstructured presentation videos is the one by Haubold et al. [7], who have conducted user studies comparing head vs. head and shoulders person representation indexes. However, such indexes were created manually and not automatically.

2. FACE TRACKS GENERATION

Face tracks are sequences of consecutive frames in which a face is tracked. Since the videos we investigate are unedited and unstructured, we cannot rely on standard shot detection algorithms to segment the video into shots. We therefore implemented a simple loose shot boundaries detector, which splits each frame into 9x5 regions, and then thresholds mean gray scale differences between corresponding regions in frames separated by a step 3 frames. While simple, this loose shot boundaries detection algorithm is 98% accurate and prevents inconvenient behaviors of the tracker.

In order to find "seed" faces (where to initialize the tracker), we use the Viola Jones face detector. To alleviate the significant amount of false detections originating from the noisy videos, we applied the skin color filter introduced by Gomez et al. [6] to each pixel in the candidate face regions. The resulting skin model in the RGB colorspace is the following:

$$Pixel = skin \iff \begin{cases} R/G > 1.185 & and \\ \frac{R*B}{(R+G+B)^2} > 0.107 & and \\ \frac{R*G}{(R+G+B)^2} > 0.112 \end{cases} \quad (1)$$

We then empirically evaluated that restricting a face track seed to require that more than 20% of the pixels in a candidate region had skin tone, resulted in doubling face track detection precision performances with respect to the default

*Area chair: Bernard Merialdo

This research was supported in part by NSF grant IIS-0713064.

MM'11, November 28–December 1, 2011, Scottsdale, Arizona, USA.

Video	#Frames	#Speak.	#GTracks	#DTracks	GATL	DATL	TPrec.	TRecall	TF1	Radius
1	85K	19	77	213	1718	830	0.896	0.637	0.745	7.11
2	103K	19	77	378	2249	640	0.916	0.584	0.713	6.01
3	65K	20	72	204	1367	551	0.921	0.581	0.713	5.92
Total	253K	58	226	795	1787	673	0.911	0.6	0.723	5.34

Table 1: Experiments videos ground truth information and tracking performances: number of frames, number of speakers, number of ground truth (G) and detected (D) tracks and Average Track Length (ATL, in frames). Tracking performances in terms of Precision, Recall, F1 and average Euclidean distance(Radius) between ground truth region and system region.

Viola Jones face detector, while maintaining the same level of recall. Once the seed has been established for a face track, we track the face in both temporal directions, until the track exits the frame borders or one of the detected shots boundaries is encountered.

We use the appearance based online multiple instance learning tracker (MILTrack) recently introduced by Babenko et al. [2], integrating it with face detections to alleviate the drifting effect in the following way. We start at a track seed, and initialize MILTrack with it. At each frame t, if we are within the loose shot boundaries determined as described above and the predicted position of the tracked region is not outside the frame, we proceed with the tracking step.

In case the Viola Jones detector finds a region $\mathbf{f}_t^O$ overlapping the output of MILTrack $\mathbf{f}_t^P = (x, y, w)$, we then consider $\mathbf{f}_t^P$ to be the noisy prediction part of a simplified, steady-state Kalman filter, while the Viola Jones detection provides a noisy observation $\mathbf{f}_t^O$. A parameter α determines how to optimally weight the prediction over the observation, and it is fixed a priori between 0 and 1. Therefore at each frame t the position of the tracked face will be fixed according to the following Equation:

$$\mathbf{f}_t \leftarrow \alpha \mathbf{f}_t^P + (1-\alpha)\mathbf{f}_t^O \quad (2)$$

The tracking process is re-initialized in case another face track seed $\mathbf{S}_t$ is encountered while tracking, since the confidence of being correctly on target in a track seed is extremely high. In Table 1 we report the tracking precision, recall and F1, as well as the average L2 distance between ground truth and tracked region face centers, obtained with the best $\alpha = 0.6, 0.5, 0.8$ for videos 1, 2 and 3 respectively.

3. OPTIMAL FACE SELECTION

We propose to process face tracks to select the most useful faces for both track matching and indexing purposes. We perform such selection based on the following three quality measures [5, 8].

Pose. We select faces presenting a 3/4 view, following the literature [3, 8] and the results of an informal user study in which we asked people which pose representation of a face they preferred to be shown to them as part of a visual face index. In order to do so, we trained a left 3/4 and right 3/4 pose detector using 1200 images from the FaceTracer[1] dataset. Each classifier is an SVM with RBF kernel based on edge histogram extracted in 5x5 uniformly split regions in an image.

SkinRatio. We select faces with a high fraction of the image occupied by skin pixels, using the filter introduced in Equation 1. This measure is useful to exclude samples where the tracker drifted away from the face of a speaker.

[1]http://www.cs.columbia.edu/CAVE/projects/face_search

Resolution. We select faces that are large. The low resolution of the videos in our dataset (in particular video 3, 432x240) demands that the index must contain face images with as much close-up as possible.

We want to stress that our approach for face selection within a track is independent from the matching descriptor choice across tracks. In order to perform across tracks matches, we chose to represent each face with the Local Binary Pattern (LBP) descriptor [1]. We split each face into a 7x7 grid and concatenate LBP histograms computed from all the regions into a 2891 dimensional feature vector $\mathbf{v}$. Finally, we use the square root of the Euclidean distance between feature vectors as a metric to evaluate the similarity between two feature vectors $\mathbf{v1}$ and $\mathbf{v2}$.

In order to select faces for the final speakers visual index, we took into account another result of our informal user study: people preferred to be shown a head-and-shoulder representation of a speaker, rather than simply the face by itself. Therefore, while the quality analysis was conducted on the face region, when producing the visual indexes reported in Figure 3 we enlarged the face bounding box by 20% horizontally and vertically, and further duplicated its height to include a head and shoulder view.

4. EXPERIMENTS

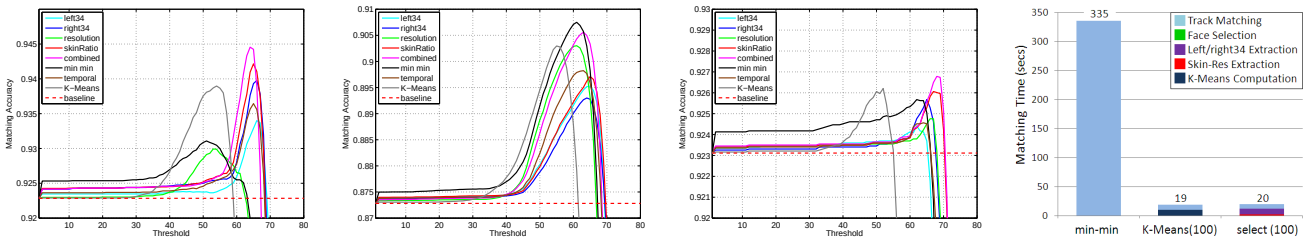
We ran experiments on 3 different MPEG videos containing student presentations. Each video is approximately 45 minutes long, for a total of more than 2 hours of footage and one quarter of a million frames. The videos were recorded by non-professionals and are unedited. They also present challenges in that the camera is rarely steady, there are not clean cuts to identify shots, resolution is low, and they lack structure. Table 1 presents the information about number of speakers and tracks of the inspected videos.

4.1 Face Tracks Extraction and Matching

In order to perform matching between tracks, a standard approach used among others by Everingham et al. [4] consists in computing the min-min distance between tracks T1 and T2, that is, compute the distance between each possible pair of elements $t1 \in T1$ and $t2 \in T2$.

We tested two methods to selection subsets of elements to match in each track: the first one involves a simple temporal sampling of the faces in the track, while the second consists in an unsupervised method based on image quality measures. For temporal sampling, we simply selected one face every n examples, and we tried values of $n = 1, 3, 10, 100$.

We evaluated track matching accuracy for each video, using the extracted tracks with best tracking precision in each video. Matching is performed in a Nearest Neighbor classification framework: given a reference track T_r, we compute



(a) (b) (c) (d)

Figure 1: (a)-(c) Matching accuracy performance for the three investigated videos. (d) Average processing time (in seconds) for track matching. Comparison between the min-min standard approach, K-means clustering (in dark blue) and the proposed selection method, which is based of 4 steps: skinRatio and image resolution extraction (2.46 seconds, in red), pose classifier evaluation (9.08 seconds, in violet), face selection (0.02 seconds, in green) and track matching (8.18 seconds when the top 100 faces for each track are retained, in light blue). Note that, unlike our proposed method, K-means does not provide face indexes.

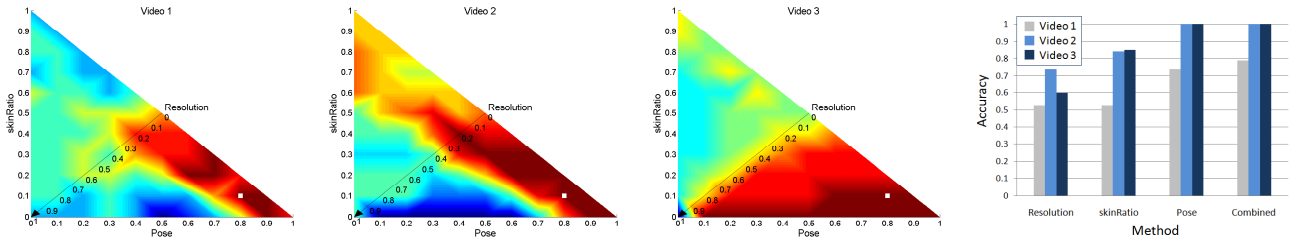


Figure 2: Selection accuracy for index building on the three investigated videos. Heat map of the accuracy given combinations of quality measures in Equation 3. The white squares represent the optimal combination, which is $\mathbf{w} = (0.8, 0.1, 0.1)^T$, interestingly shared across all three videos.

the distance between the reference and all the tracks T_i in the video which do not temporally overlap with it (we consider that if two tracks overlap temporally, they must belong to different individuals). We retain the track T_i which has the smallest distance $d(r, i)$ to T_r to be a candidate match, and if $d(r, i)$ is smaller than a matching threshold, we consider that T_r and T_i are a match. We compared four different modalities to compute the distance between two tracks.

The first modality (min-min) is to compute the distance between all pairs of images $d(\mathbf{v}_r^m, \mathbf{v}_i^n)$, where $\mathbf{v}_r^m \in T_r$ and $\mathbf{v}_i^n \in T_i$, and keep the minimum distance to represent the distance between the tracks. This method presents two limitations in our unconstrained domain. The first one is efficiency, given the large amount of redundancy present in each track due to the temporal domain of a video. The second is accuracy, since the face tracks generation module returns a set of tracks which can be considered noisy, containing false positives from the seeds detection stage, and drifts in tracking caused cutting of face regions.

The second modality involves using a temporal sampling of n faces in each track, computing the distances between these reduced sets and retain the minimum one.

The third modality employs K-means clustering in the LBP feature space, as suggested by Mau et al.[9]. The $K = n$ cluster centers are used to compute the distances between tracks, and the minimum distance is kept.

The last modality consists in computing the distances only between the n top track faces which are selected based on the response to our quality measure filters (see Section 3).

For temporal, K-means, and selection matching we retained $n = 100$ samples from each track (or the whole track in case the length of the track is smaller than 100). Experiments with $n = 1, 3, 10$ were also conducted and provided worse matching accuracy results, which we omit for brevity.

Figure 1 reports track matching accuracy as the threshold varies in percentage of the range of values of the track distances. In the Figure we report a baseline which predicts all pairs to be non-matches. For all videos, all selection-based matching methods present a global maximum in the accuracy with respect to the chosen threshold. The optimal threshold seems to be consistently located between 60% and 70% of the range of distances between tracks. From the results shown in Figure 1, not only the computational cost of computing the distances between tracks is reduced when using filtering techniques to reduce the number of image pairs to match, but matching accuracy increases in two out of three videos. This is due to the reduction or removal of noisy, drifted, or partially cut images from the tracks which is accomplished through sampling. It is also interesting to notice that the best filter results in such videos is the skin color-based one. In fact, proper face matching requires a full face occupying most of the image, which is best guaranteed by the skin color filter.

Figure 1(d) shows the computational gain of using the proposed face selection method within tracks before matching. Notwithstanding the overhead introduced by feature extraction and face selection before matching, the proposed approach achieves a higher level of track matching accuracy while needing approximately 6% the running time of min-min matching. This is due to the greatly reduced computational complexity of matching each pair of tracks: $O(k^2)$ (with $k = 100$ in our experiments) versus $O(n^2)$ for min-min, where n is the number of frames in a track. According to Table 1 the relationship between k and n is on average of 1 to 6.7, which is quite significant when squared.

We tested a series of possible combinations of the three

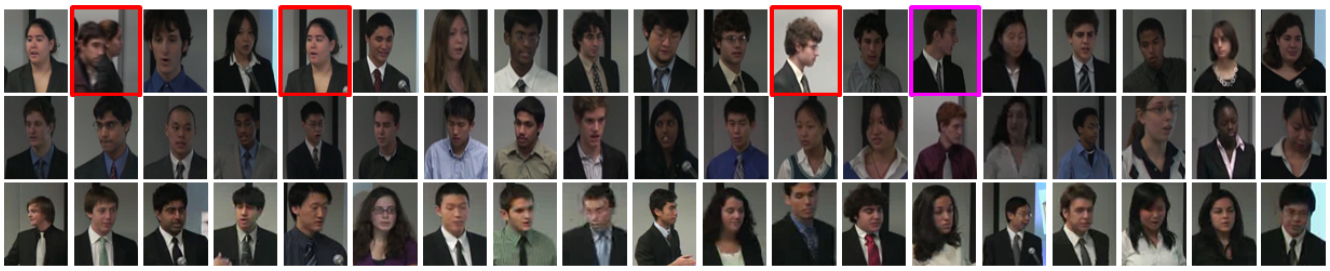


Figure 3: Generated visual speaker index. Most of the images show the desired 3/4 head and shoulder view of the speaker. Some fail, either by portraying the wrong person (in red) with respect to the ground truth or by presenting a full profile (in magenta), from which is hard to identify the person.

quality measures according to the following formula (each quality measure is normalized between 0 and 1):

$$Q = w1 \cdot pose + w2 \cdot resolution + w3 \cdot skinRatio \qquad (3)$$

with *pose* being either $left34$ or $right34$. We empirically found that the best combinations of $\mathbf{w} = (w1, w2, w3)^T$ were $(0.0, 0.3, 0.7)^T$, $(0.3, 0.1, 0.6)^T$ and $(0.0, 0.7, 0.3)^T$ for video 1, 2 and 3 respectively, and $pose = right34$ for all videos. It is interesting to notice how *resolution* is much more important for the low resolution video 3, while *pose* does not seem to be fundamental, probably because faces with different poses were matched against each other.

We also note that the performance of our selection method is comparable with K-means clustering both in terms of accuracy and efficiency. However, our selection method provides us also with the candidate faces for the speakers visual index, whereas the average faces returned by K-means do not hold any meaning to a human.

4.2 Representative Index Extraction

In order to obtain the faces to build the speakers visual index, we took the results of the 3 filters presented in Section 3 to all the images in each track and retained the ones returning the best combined scores, following Equation 3 (with the modification that $pose = max\,|left34, right34|$, since differently from matching we do not care which direction the face is facing, as long as it is a 3/4 view). The best face among all those in the tracks representing the same speaker (resulting from track matching) is expanded to include a head and shoulder view of the person.

In Figure 2 is reported the accuracy of the indexes obtained with different combinations of the three quality measures, as well as their individual performances. Accuracy is measured as the fraction (out of the possible 58 speakers) of selected images representing a 3/4, head and shoulder view of a speaker. In this framework, differently from matching, the *pose* measure is the predominant factor in performance. This is because from the perspective of the visual index, a full frontal or full profile view of a person is considered an error. It is also interesting to notice that while in video 1 and 2 *resolution* seems to hurt in combination with the other two measures, on the lowest resolution video 3 this effect is not registered (lower triangle of the heat map).

Figure 3 shows the head and shoulders views of speakers selected for the visual index. The system is able to automatically generate a qualitatively pleasing index consisting of 54 out of the 58 speakers present in the videos. One speaker was never detected, and in some cases the wrong person or view were selected since tracks were not matched properly or *resolution* and *skinRatio* prevailed over *pose*.

5. CONCLUSIONS

We have presented a system to select the most representative faces in unstructured presentation videos with respect to two criteria: to optimize matching accuracy between pairs of face tracks, and for indexing purposes.

We were able, by using quality metrics, to build face indexes of the speakers in 3 unstructured presentation videos, with a head and shoulders, 3/4 view, which is the pose preferred by humans. In the future we plan to evaluate the performance of the generated indexes through user studies.

6. REFERENCES

[1] T. Ahonen, A. Hadid, and M. Pietikainen. Face recognition with local binary patterns. In *ECCV*, pages 469–481, 2004.

[2] B. Babenko, M.-H. Yang, and S. Belongie. Visual Tracking with Online Multiple Instance Learning. In *CVPR*, 2009.

[3] D. Burke, J. Taubert, and T. Higman. Are face representations viewpoint dependent? a stereo advantage for generalizing across different views of faces. *Vision Research*, 47(16):2164 – 2169, 2007.

[4] M. Everingham, J. Sivic, and A. Zisserman. Taking the bite out of automated naming of characters in tv video. *Image and Vision Computing*, 27(5):545 – 559, 2009.

[5] A. Fourney and R. Laganiere. Constructing face image logs that are both complete and concise. In *CRV*, pages 488–494, 2007.

[6] G. Gomez and E. F. Morales. Automatic feature construction and a simple rule induction algorithm for skin detection. In *ICML Workshop on Machine Learning in Computer Vision*, pages 31–38, 2002.

[7] A. Haubold and J. R. Kender. VAST MM: multimedia browser for presentation video. In *CIVR*, pages 41–48, 2007.

[8] X. Liu, J. Rittscher, and T. Chen. Optimal pose for face recognition. In *CVPR*, pages 1439–1446, 2006.

[9] S. Mau, S. Chen, C. Sanderson, and B. Lovell. Video face matching using subset selection and clustering of probabilistic multi-region histograms. In *ICIVC*, pages 1 – 8, November 2010.

[10] K. Nasrollahi and T. Moeslund. Complete face logs for video sequences using face quality measures. *Signal Processing, IET*, 3:289 – 300, 2009.