On evaluating face tracks in movies

Alexey Ozerov, Jean-Ronan Vigouroux, Louis Chevallier, Patrick Pérez

ON EVALUATING FACE TRACKS IN MOVIES

Alexey Ozerov, Jean-Ronan Vigouroux, Louis Chevallier and Patrick Pérez

technicolor
975, avenue des Champs Blancs
F-35510 Cesson-Sévigné, France
Email: {firstname.lastname}@technicolor.com

ABSTRACT

Automatic extraction of *face tracks* is a key component of systems that analyse people in audio-visual content such as TV programs and movies. Due to the lack of properly annotated content of this type, popular algorithms for extracting face tracks have not been fully assessed in the literature. We introduce and make publicly available a new dataset, based on the full annotation of a feature movie, to help fill this gap. We show in particular that, thanks to this dataset, state-of-art tracking metrics can now be exploited to evaluate face tracks used by, e.g., automatic character naming systems. We conduct such an evaluation on different variants of a novel system that we introduce as a generalization of existing ones.

***Index Terms*—** Face detection, face tracking, movies, evaluation, ground truth annotation.

1. INTRODUCTION AND RELATED WORK

In the recent past, face-based analysis of people in complex video content such as TV shows, TV series and movies has become an important research area. It is motivated by growing needs for annotation, navigation and search tools in audiovisual archives. Tasks under concern include in particular people clustering and identification [1], and character naming [2, 3]. The most popular pre-processing step for such systems is the automatic extraction of *face tracks*, via a combination of face detection and temporal grouping at the shot level. These face tracks then become the unit of analysis: person clustering amounts to grouping tracks stemming from a single person, while person naming is the task of deciding whether a face track corresponds to one of the characters in a pre-defined list.

Quantitative evaluation of such systems is a crucial issue. For practical reasons, it is usually conducted on datasets where face tracks are *automatically* extracted by the same processing step as in evaluated system, and *manually* labeled [2]. This protocol leads to the following two limitations. First, it binds produced ground-truth to a specific face track detector and thus only allows the evaluation of pipelines that rely on this detector. Second, not relying on manually extracted face tracks makes system evaluation only partial. Performance of the final task, e.g., naming, can be assessed, but not that of the complete analysis pipeline, from detection to naming. In particular, face tracks that are missed due to, e.g., poor lighting, partial occlusion or motion blur, have no impact on performance metrics, and system's overall recall can be vastly over-estimated. Acknowledging this problem of "coverage", manually annotated face and identity ground-truths were recently created [3, 4]. However, this annotation remains partial: it is not temporally dense (every I-frame in [4] and every 10th frame in [3]) and only faces of main characters were annotated in [3].

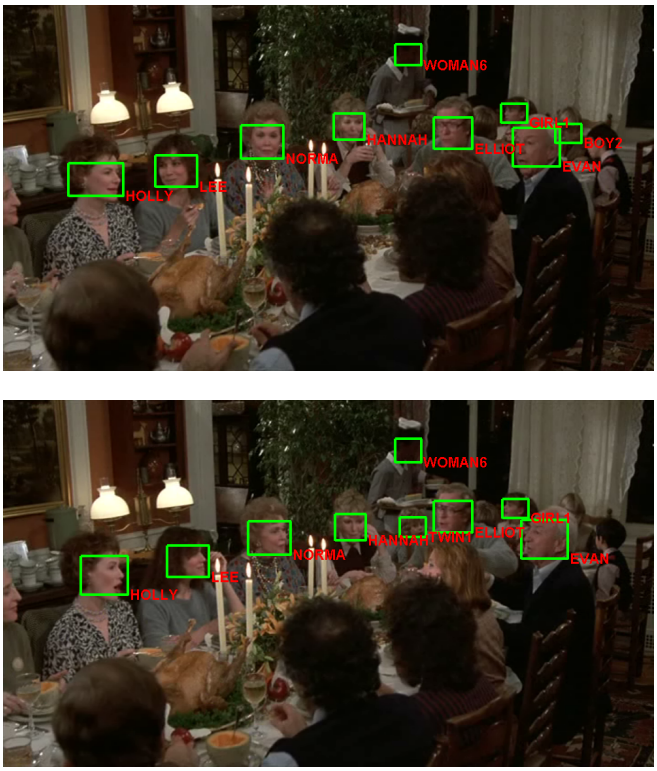

Fig. 1: **Example of spatio-temporal face annotation in *Hannah* dataset**: two frames, 1.5s apart, from a shot in "Hannah and her sisters"; visible faces are identified with labeled bounding boxes.

To overcome the above shortcomings, we introduce a new dataset based on the dense annotation of a complete feature movie of 1 hour and 40 minutes. Among other annotated information, more than 2,002 face tracks were manually extracted from 153,833 frames and identified (see example in Fig. 1). To the best of our knowledge, it is the first time movie data are annotated for faces at such a scale and such a level of density. We believe it will be a valuable tool for the community when evaluating automatic person annotation pipelines, whether as a whole, or in parts.

This work was partially funded by the QUAERO project supported by OSEO and by the European integrated project AXES.

In particular, having access to manual ground-truth at the face track level allows us to deploy modern detection and tracking evaluation metrics [5, 6]. We demonstrate the use of such metrics, including for instance "track purity", on several variants of a new system that combines different state-of-art face detectors [7, 8].

This system is also a contribution of our work. Strongly related to popular approaches proposed in [2, 3, 9], it unifies and extends them in a sense that will be highlighted below.

In summary, the contribution of this work is the following: (1) Introduction of a novel dataset with face tracks annotated over the full extent of a feature movie; (2) Introduction of a unified approach for offline face track extraction that merges the results of different face detectors; (3) Evaluation and comparison of several variants of this approach. Two first variants are very similar to state-of-the-art approaches such as [2, 3], which are thus evaluated for the first time against fully manual face track annotation.

2. THE *Hannah* DATA SET

The dataset we introduce here, named *Hannah*, is based on the movie "Hannah and her sisters" by Woody Allen, released in 1986 and available on DVD. The full movie has been manually annotated for several types of audio and visual information. Although this is not the focus of present paper, let us mention that audio tracks have been manually annotated with speech segments and associated speaker identification. On the visual side, annotation concerns all shot boundaries and all identified face tracks within shots. This visual annotation work has been achieved using the VIPER-GT platform (`http://viper-toolkit.sourceforge.net/`). Audio annotation was performed using Audacity (`http://audacity.sourceforge.net/`). Complete audio-visual annotations are publicly available at `https://research.technicolor.com/rennes/hannah/`.

The face ground-truth metadata contains a frame by frame description of all "sufficiently" visible faces in the form of a horizontal, rectangular bounding box and an identifier. All the annotations were performed by a single annotator who was given the following instructions: all the poses from frontal to profile are accepted; for a face to be annotated, corresponding bounding box should be wider than 24 pixels (image size being 996×560); bounding box goes vertically from the middle of the chin to the middle of the forehead and, horizontally, from one ear to the other or from one ear to the tip of the nose depending on the pose (see Fig. 2 for examples); finally, regarding occlusion, it was required that at least half of the face was visible.

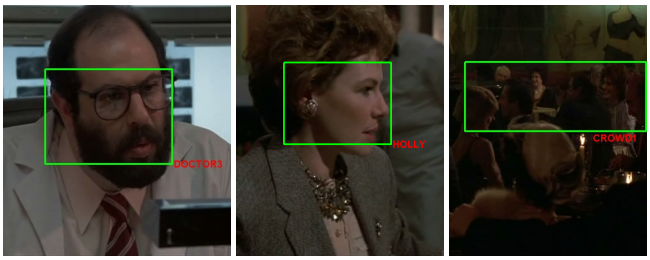

Fig. 2: **Annotated bounding boxes**: faces that are sufficiently large and visible are singled out by an individual box, following precise annotation specifications; For groups of hardly distinguishable faces, a collective "crowd" box annotation is used. "Crowd" boxes are ignored in the evaluations.

Each bounding box is also manually tagged, based on the identity of the person. For 53 main characters, the label is the name, such as "Hannah", "Elliot" and "Lee" in Fig. 1. For other persons, 186 of them have been uniquely identified and tagged with labels such as "Girl1" or "Boy2" in Fig. 1. Finally, in crowded scenes, groups of secondary characters were annotated within collective bounding boxes and labeled as 'Crowd1', 'Crowd2', etc. (Fig. 2). There is a total of 254 distinct labels in the dataset.

Given face and shot annotations, *ground-truth face tracks* ("GT-track" in short) are defined as follows: a face track is a maximally long sequence of face bounding boxes that are consecutive in time, share the same label and belong to the same shot. There are 2,002 such tracks spread over the 245 shots of the movie. Duration of GT-tracks ranges from 1 to 500 frames, with a mean of 99.1 frames. The number of tracks simultaneously appearing in a frame ranges from 0 to 10 and more in the numerous gathering scenes.

3. A UNIFIED EXTRACTION OF FACE TRACKS

Following [2, 3, 9], our approach is based on the following main steps: (1) Shot boundaries are automatically detected; (2) Faces are detected within each frame, using one or several face detectors; (3) Faces detected in a shot are grouped in time through bi-directional matching/tracking of local features; (4) Face groups are turned into final face tracks through temporal interpolation and smoothing. Below, we present these steps in details and discuss their connection to the state-of-art approaches from which they derive.

Shot boundary detection. Classically, shots are identified through the detection of their boundaries. A large choice of such detectors is available [10]. Our pipeline relies on the one proposed by [11] in the context of video fingerprinting. It is based on monitoring the temporal evolution of so-called RASH image descriptors [12, 13].

We have evaluated this detector against the shot ground-truth in *Hannah*, using the metrics in [10]. Defining as true positive a detected shot boundary that falls at no more than e frames from a true one ($e = 1$ in our experiments), precision, recall and F measures are classically computed. On the 244 cuts of the dataset, precision, recall and F measures obtained with the RASH-based detector are respectively 0.934, 0.808 and 0.866. Its performance is on the par with state-of-art shot boundary detectors, while processing is 10 times faster than real time.

Image-wise face detection by multiple detectors. Given a set of different face detectors, each of them is applied to each image within the current shot. As a result, N bounding boxes, B_i, $i = 1 \cdots N$ are output for this shot. The i-th bounding box B_i is defined by its time stamp t_i and by its location $\mathbf{p}_i = (x_i, y_i, w_i, h_i)$, where (x_i, y_i) are its center coordinates, w_i its width and h_i its height.

In our implementation we consider up to four different detectors. First, we consider three instances of Viola and Jones (VJ) face detectors [7] trained, respectively, for frontal, left profile and right profile. Second, we consider a recent detector proposed by Zhu and Ramanan (ZR) [8] that can handle multiple poses by itself.

Grouping detections. The N detections in the shot are grouped by a spatio-temporal clustering which performs both bi-directional face tracking and instantaneous grouping of multiple detections. To this end, we define a similarity measure between two detected faces B_i and B_j that depends on their temporal separation $t_i - t_j$ and can take the particular form of a *cannot link* (CL) constraint in case they are in same frame ($t_i = t_j$). Let's first define the asymmetric similarity a_{ij} as follows:

- If $t_i = t_j$, $a_{ij} = \frac{|B_i \cap B_j|}{|B_i|}$.

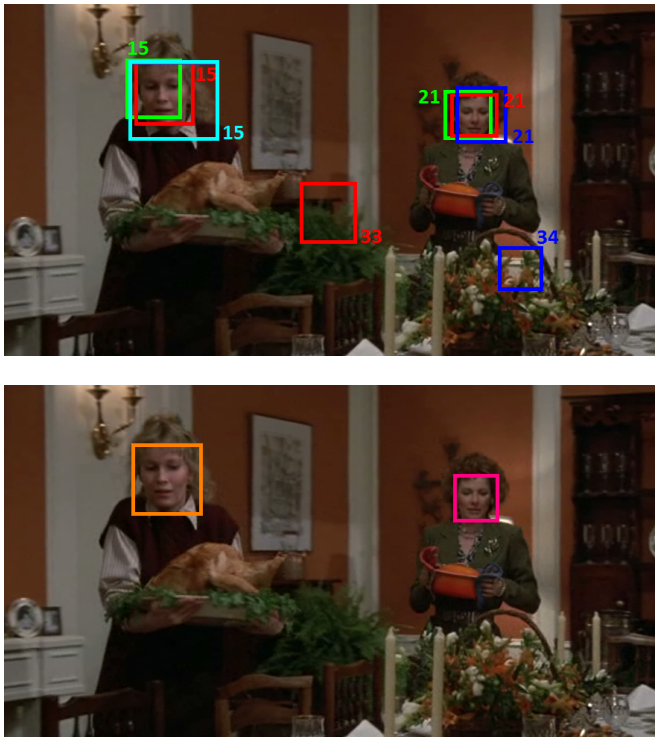

Fig. 3: **From multiple face detections to face tracks**: (Top) All faces detected in a given shot are grouped by spatio-temporal clustering. In some cases, outputs of different detectors in a same frame (colors of bounding boxes correspond to four different detectors) can end up in a same cluster, as visible from cluster IDs. (Bottom) Face clusters are turned into face tracks of sufficient length by merging, smoothing and interpolation. Two tracks remain at this instant, out of four extracted clusters.

- If $0 < |t_i - t_j| \leq L$ (we considered $L = 5$), key points extracted in region B_i of frame t_i are "tracked" with one step of KLT tracker [14] in frame t_j. Then, a_{ij} is set as the fraction of these points whose found correspondent falls in B_j.
- If $|t_i - t_j| > L$, $a_{ij} = 0$.

The symmetric similarity s_{ij} is then defined as $s_{ij} = \min(a_{ij}, a_{ji})$. Furthermore, symmetric similarity s_{ij} is replaced by CL constraint if $t_i = t_j$ and $\max(a_{ij}, a_{ji}) > S_{CL}$ (we considered $S_{CL} = 0.25$). This relies on the hypothesis that two different detections in the same frame that have too small an overlap cannot correspond to the same person.

Based on defined similarity measure, grouping of face detections in the shot is obtained by bottom-up agglomerative clustering with single linkage and CL pair-wise constraints [15]. A given cluster usually spans several time instants, which are not necessarily consecutive. Also, it can exhibit more than one detection for a given instant. One such example of multiple detections grouped in the same frame is given on Figure 3, top.

Face tracks creation. Once the face detections are clustered, the creation of final face tracks requires (i) to merge simultaneous detections of same faces (i.e., replace them by a single bounding box), (ii) to interpolate missed detections and (iii) to smooth bounding boxes over time. Inspired by [9], we formulate these three operations within the following single optimization problem, for a given cluster:

$$\widehat{\mathbf{Q}} = \arg\min_{\mathbf{Q}} \sum_{i \in I} \omega_i^2 \|\mathbf{q}_{t_i} - \mathbf{p}_i\|^2 + \lambda^2 \sum_{t=t_{min}}^{t_{max}-1} \|\mathbf{q}_t - \mathbf{q}_{t+1}\|^2, \quad (1)$$

where $\mathbf{p}_i$ defines the i-th detected bounding box in the shot, I is the index subset for detections in the cluster, $\mathbf{q}_t = (x_t, y_t, w_t, h_t)$ defines a unique bounding box to be extracted in frame t and unknown face track is $\mathbf{Q} = \{\mathbf{q}_t\}_{t=t_{min}\cdots t_{max}}$ with t_{min} and t_{max} the minimum and maximum time stamps in $\{t_i, i \in I\}$. Parameter λ (set to $\lambda = 8$ for 25 Hz videos) controls temporal smoothing, while parameters ω_i's permit to balance the importance of the different detectors if needed (we have chosen $\omega_i = 0.5$ for VJ profile detectors and $\omega_i = 1$ for all other detectors). Optimizing (1) is equivalent to solving a tri-diagonal linear equation. This is done efficiently, with a linear complexity with respect to the track length, by means of Gaussian elimination with partial pivoting [9].

As a final step, face tracks shorter than M frames (we considered $M = 16$ for 25 Hz videos) are removed. See Figure 3 for an example of the face track creation procedure.

Relation to existing methods. As was already mentioned, our approach mostly follows those proposed in [2, 3, 9]. However, there are some small differences that makes it appealing. First, while the idea of merging profile and frontal Viola and Jones detections was already used in [3], here we extended it to allow the joint exploitation of completely different face detectors. As an example which is evaluated in next section, we seamlessly combine Viola and Jones detectors with detector recently proposed in [8]. Second, in [2, 3] KLT tracking is performed through adjacent frames, whereas we here use KLT tracking between frames lying within L frames distance. This means our tracker can handle occlusions that last up to $L-1$ frames. Also, in contrast to [2, 3], our agglomerative clustering takes into account additional pair-wise cannot-link constraints to avoid inappropriate groupings of simultaneous detections. Finally, we formulate merging, interpolation and smoothing of clustered detections as a single optimization problem while, in [2, 3, 9], these operations are performed sequentially.

4. EVALUATION USING *Hannah*

4.1. Metrics for detections and tracks

Automatic multi-object tracking systems can be evaluated in two complementary ways: at the frame level with detection-based metrics and at the temporal level with track-based metrics [5,6]. Thanks to dense spatio-temporal annotation in *Hannah* dataset, we can conduct both types of performance assessment.

At the frame level, the configuration of boxes from *extracted face tracks* ("e-track" in short) has to be compared with those from GT-tracks. Let G one of the GT-track boxes in the frame and E, one of the e-track boxes in the same frame. These two boxes are declared a "match" if they overlap sufficiently, that is, if their F-measure $\frac{2\rho\nu}{\rho+\nu}$, where $\rho = \frac{|G \cap E|}{|G|}$ and $\nu = \frac{|GT \cap E|}{|E|}$, is above a threshold t_C, set to 0.33 in our evaluations, as suggested in [6]. A binary matching matrix can be built, with one row per GT-track in the frame and one column per e-track. Based on it, evaluation at the frame level can be conducted by counting following errors [5]:

- *False Positive* (FP): an e-track box with no match.
- *False Negative* (FN): a GT-track box with no match.
- *Multiple Track* (MT): a GT-track box with multiple matches.

For a time instant, these quantities are normalized with respect to the total number of GT-track boxes in the frame. Averaging over all frames provides average quantities $\overline{\text{FP}}$, $\overline{\text{FN}}$ and $\overline{\text{MT}}$.

Moving to track-based metrics requires to monitor matching between ground-truth tracks and extracted tracks over time [6]: a GT-track is said to be identified by the e-track to which it is associated the most often along its timespan and its purity (*object purity*, OP) is the corresponding fraction of time; similarly, an e-track is said to identify the GT-track to which it is associated the most often along its timespan and corresponding time fraction is its purity (*tracker purity*, TP). Both purities are averaged over all GT-tracks and e-tracks respectively, yielding quality measures $\overline{\text{OP}}$ and $\overline{\text{TP}}$. Finally, purity defined as

$$P = 2\frac{\overline{\text{TP}}\,\overline{\text{OP}}}{\overline{\text{TP}}+\overline{\text{OP}}}, \tag{2}$$

measures the overall quality of all face tracks.

4.2. Results

The face track extraction system that we introduced in section 3 can jointly accommodate any types of face detectors. Choosing different combinations of face detctors, we build four versions of the system: "Front.", "Front.+Prof.", "ZR" and "All" = "Front.+Prof.+ZR", where "Front." stands for VJ frontal detector, "Prof." stands for two VJ profile detectors (left and right)[1] and "ZR" stands for Zhu and Ramanan detector.[2] We also provide an evaluation of raw (multiple) face detections with frame-based metrics only, so as to assess the ability of face track extractor to merge multiple detections, filter-out false detections and interpolate missed detections.

Some basic statistics on the ground truth and the systems outcomes are summarized in table 1, where #boxes stands for the number of boxes, #tracks stands for the number of tracks, and MTL stands for the mean track length.

	Tracking	#boxes	#tracks	MTL
Ground-Truth	-	198224	2002	99.1
Front.	No	94395	-	-
Front.+Prof.	No	229991	-	-
ZR	No	85663	-	-
All	No	315654	-	-
Front.	Yes	63050	1204	52.4
Front.+Prof.	Yes	118090	790	149.5
ZR	Yes	72356	1057	68.5
All	Yes	136904	847	160.7

Table 1: Basic statistics on the ground truth and the systems outcomes.

As expected, combining several detectors leads to more raw detections and to fewer tracks that are longer in average. Interestingly, tracks based on multiple detectors are longer in average than those of the GT. We explain this by the fact that while our approach allows tracking through some partial and brief total occlusions, in the GT such occlusions may lead to track splitting (see the annotation specifications in section 2).

[1] As implemented in OpenCV 2.4.3.

[2] Source code provided by the authors, parameters set as in reference article (threshold $= -0.65$), using intermediate model Face_small_146filters.xml

Quality measures presented earlier allow us to compare the four versions of the system, both in terms of instantaneous detection errors and of face track quality, and the results are given (in percent) in tables 2 and 3, respectively.

	Tracking	$\overline{\text{FP}}$ (%)	$\overline{\text{FN}}$ (%)	$\overline{\text{MT}}$ (%)
Front.	No	22.9	61.3	0.179
Front.+Prof.	No	49.7	40.3	35.600
ZR	No	6.6	54.0	0.005
All	No	55.4	32.7	60.000
Front.	Yes	8.0	67.8	0.000
Front.+Prof.	Yes	15.8	47.0	0.165
ZR	Yes	2.1	58.3	0.002
All	Yes	17.4	39.2	0.390

Table 2: Evaluation of four tracking systems as well as corresponding raw detections in terms of instantaneous detection.

	Tracking	$\overline{\text{OP}}$ (%)	$\overline{\text{TP}}$ (%)	Purity (%)
Front.	Yes	8.2	73.4	14.7
Front.+Prof.	Yes	17.9	47.0	26.1
ZR	Yes	11.9	91.5	21.0
All	Yes	22.5	50.6	31.2

Table 3: Evaluation of four tracking systems in terms of purity.

Let us first comment on raw face detection results (without tracking) from table 2. Naturally, while combining detectors allows reducing $\overline{\text{FN}}$, it leads at the same time to excessively high $\overline{\text{FP}}$ and $\overline{\text{MT}}$. Tracking allows considerably reducing $\overline{\text{FP}}$ and $\overline{\text{MT}}$ at the expense of a moderate increase in $\overline{\text{FN}}$, possibly due to filtering-out very short tracks that can be sometimes correct. Finally, the best tracking results in terms of purity (see table 3) are achieved by the system combining the four detectors, thus confirming a potential of our approach. Note however that overall the results in terms of detection seem to be poor, as compared to the results obtained by corresponding face detectors on standard databases of still images [7, 8]. This indicates a very challenging nature of *Hannah* dataset, which contains very small faces, a wide range of poses, difficult lighting conditions and cluttered scenes.

5. CONCLUSION

We have created and made freely available *Hannah* dataset that should help the community assessing existing and future systems for face detection, tracking, identification and naming, as well as for speaker diarization. We have also proposed an original face track extraction system that can easily benefit from a number of different detectors. Evaluated on *Hannah* dataset in its various configurations, it was shown that thanks to its extensibility the system is able to take advantage of four detectors, which provides superior performance, as compared to using fewer detectors. Regarding evaluation in terms of detection the number of false negatives is about 40% which is above what is published on face detection performance for still images. To our knowledge this is the first time evaluation of face detection and tracking is performed on a full movie densely and manually annotated.

6. REFERENCES

[1] R. G. Cinbis, J. Verbeek, and C. Schmid, "Unsupervised metric learning for face identification in TV video," in *International Conference on Computer Vision*, Barcelona, Spain, Nov. 2011.

[2] M. Everingham, J. Sivic, and A. Zisserman, "Taking the bite out of automatic naming of characters in TV video," *Image and Vision Computing*, vol. 27, no. 5, pp. 545–559, 2009.

[3] J. Sivic, M. Everingham, and A. Zisserman, ""Who are you?" – learning person specific classifiers from video," in *Proc. CVPR*, 2009.

[4] R. Kasturi, D. Goldgof, P. Soundararajan, V. Manohar, J. Garofolo, R. Bowers, M. Boonstra, V. Korzhova, and J. Zhang, "Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 319–336, Feb. 2009.

[5] K. Smith, D. Gatica-Perez, J.-M. Odobez, and S. Ba, "Evaluating multi-object tracking," in *In Workshop on Empirical Evaluation Methods in Computer Vision*, 2005.

[6] K. Smith, "Bayesian methods for visual multi-object tracking with applications to human activity recognition." Ph.D. dissertation, École Polytechnique Fédérale de Lausanne, Feb. 2007.

[7] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. CVPR*, 2001, pp. 511–518.

[8] X. Zhu and D. Ramanan, "Face detection, pose estimation and landmark localization in the wild," in *Proc. CVPR*, June 2012.

[9] A. Kläser, M. Marszalek, C. Schmid, and A. Zisserman, "Human focused action localization in video," in *SGA*, 2010.

[10] C. Cotsaces, N. Nikolaidis, and I. Pitas, "Video shot detection and condensed representation. a review," *Signal Processing Magazine, IEEE*, vol. 23, no. 2, pp. 28–37, Mar. 2006.

[11] A. Massoudi, F. Lefèbvre, C.-H. Demarty, L. Oisel, and B. Chupeau, "A video fingerprint based on visual digest and local fingerprints," in *Image Processing, 2006 IEEE International Conference on*, Oct. 2006, pp. 2297–2300.

[12] F. Lefèbvre, B. Macq, and J.-D. Legat, "RASH: RAdon Soft Hash algorithm," in *11^{1}h European Signal Processing Conference*, Toulouse, France, Sep. 2002.

[13] C. De Roover, C. De Vleeschouwer, F. Lefèbvre, and B. Macq, "Robust video hashing based on radial projections of key frames," *Signal Processing, IEEE Transactions on*, vol. 53, no. 10, pp. 4020–4037, Oct. 2005.

[14] J. Shi and C. Tomasi, "Good features to track," in *Proc. CVPR*, 1994, pp. 593–600.

[15] S. Miyamoto and A. Terami, "Semi-supervised agglomerative hierarchical clustering algorithms with pairwise constraints," in *Proc. FUZZ'10*, 2010, pp. 1–6.