A Complete Framework for Fully-automatic People Indexing in Generic Videos

Dario Cazzato[1], Marco Leo[2] and Cosimo Distante[2]
[1]*Department of Innovation Engineering, University of Salento, Lecce, Italy*
[2]*National Research Council of Italy - Institute of Optics, Arnesano (Lecce), Italy*

Keywords: Hartigan Index, Silhouette, Face Indexing, People Identification, Clustering.

Abstract: Face indexing is a very popular research topic and it has been investigated over the last 10 years. It can be used for a wide range of applications such as automatic video content analysis, data mining, video annotation and labeling, etc. In this work a fully automated framework that can detect how many people are present in a generic video (even having low resolution and/or taken from a mobile camera) is presented. It also extracts the intervals of frames in which each person appears. The main contributions of the proposed work are that no initializations neither a priory knowledge about the scene contents are required. Moreover, this approach introduces a generalized version of the *k*-means method that, through different statistical indices, automatically determines the number of people in the scene.

1 INTRODUCTION

Today videos represent one of the most important media. Every Internet user can upload his own videos and avail of what others create, and in this way video sharing has become increasingly habitual among web users. Since this huge amount of material is growing, new efficient techniques of automatic video annotation had been investigated (Hu et al., 2011). Face indexing is the technique to automatically label faces in a scene: this is a very challenging problem, due to the high variation of the pose, facial expressions and lighting conditions for a person in the same video, but it allows a lot of applications like TV shows video analysis, automatic labeling of characters in a movie (Delezoide et al., 2011) or to improve image retrieval process on a huge amount of data (Hao and Kamata, 2011).

A first attempt to obtain automatic labeling of people in a movie using speech and face recognition techniques was given by (Satoh et al., 1999), where names and faces in news videos were associated by using face recognition techniques combined with methods that extract candidate labels from transcripts. In (Pham et al., 2008) the asymmetry between visual and textual modalities was exploited, building a cross-media model for each person in an unsupervised manner, dealing with the fact that the number of faces can be different from the number of names. In the work of (Sivic et al., 2009) facial features were tracked over time and facial descriptors invariant to pose were defined. This way the authors empowered the recognition task and created a framework to label characters in TV series.

Automatically annotation of faces in personal videos by combining the grouped faces by a clustering method with a weighted feature fusion was presented in (Choi et al., 2010). The method dealt also with color information, but it needed a training set in order to perform a general-learning (GL) training scheme. In the context of photo album, (Zhu et al., 2011) used a Rank-Order distance based clustering algorithm to groups all faces without knowing the number of clusters. (Foucher and Gagnon, 2007) grouped faces using a spectral clustering approach. A tracking algorithm was also proposed in order to form trajectories. However, in this application the choice of the optimal number of clusters was not critical, and the usage of spectral clustering implies anyway specifying the cluster number. (Arandjelovic and Cipolla, 2006) tried to automatically determine the cast of a feature-lenght film using facial information and working in a manifold space. In (Prinosil, 2011) blind separation, i.e. labeling with lack of any prior knowledge, was proposed. A face model for each face was created and compared with a similarity index. The method worked well only in case of limited number of participants, relative stable video scene and face images

Cazzato D., Leo M. and Distante C..
A Complete Framework for Fully-automatic People Indexing in Generic Videos.
DOI: 10.5220/0004653502480255
In *Proceedings of the 9th International Conference on Computer Vision Theory and Applications* (VISAPP-2014), pages 248-255
ISBN: 978-989-758-004-8

captured in frontal-view.

Summing up, from the review of the related works, it is possible to conclude that some static assumptions are always needed: a prior knowledge of the scene, a minimum quality concerning the facial images and the input videos, or a known number of the total people (often working only with a minimum number of two people). Furthermore, they don't reconstruct all the interval of appearance for each different person and their performances suffer from the variety of lighting conditions, scale and pose.

The goal of this paper is to overcome the limits of existing approaches presenting a framework to automatically obtain face indexing in a generic video. The term "generic" here means that the number of people in the scene is not a priori known, each person may appear one or more time, image data can be both of good or bad quality (i.e. acquired from high definition device, webcam or smartphone) and that images can be taken from still or mobile devices. Given a generic video as input, the outcomes of the proposed framework are not only the list of the persons in the scene (if any), but also the intervals of frames (segments) in which each person appears. To do this, a multistep framework is introduced: all facial images are, at first, extracted by the Viola-Jones face detector (Viola and Jones, 2001) and then, each face patch is vectorized and represented in a new vectorial space defined by the Principal Component Analysis (*PCA*) (Wold et al., 1987) that reduces data dimensionality. Finally, a generalized usage of the well-known k-means method (MacQueen et al., 1967) is exploited for face clustering. The most interesting aspect of this step is that the number k of clusters is not a priori known (as the classical k-means requires). The idea to find the best value of an unknown k could be faced by Dirichlet Process Mixture, but in our context it is difficult to the decide on the base distribution since the model performance will depend on its parametric form, even if defined in a hierarchical manner for robustness (Görür and Rasmussen, 2010). When this value is not known, Correlation Clustering (Bansal et al., 2004) can be used. This method finds the optimal number of clusters basing on the similarity between the data points. Anyway, this approach was shown to be NP-Complete, so only approximation algorithms can be used. Also Hierarchical Clustering (Johnson, 1967) could be used, but it was discarded considering that it is effective only on small amount of data, i.e. when patterns and relationships between clusters are easily discernible. In the proposed work, the number of cluster is automatically estimated through the computation of several statistical indices after different run of the k-means algorithm with different values of k.

Summing up, the main contributions of this paper are the followings:

1. a framework that works in completely blind scenarios, i.e. where no prior knowledge is available and also in challenging scenarios, i.e. in presence of very wide range of lighting, scale, pose and facial expressions is proposed;
2. a considerable indexing precision is guaranteed even if the framework operates without supporting techniques like face tracking;
3. any video quality, taken from both still or mobile devices, also in presence of fast camera movements and noise due to shaking or facial occlusions, can be handled;
4. a generalized version of the well-known k-means method that is able to automatically determine the best configuration of the clusters embedded in the data is introduced;
5. also videos containing just one or no persons are handled.

2 OVERVIEW OF THE PROPOSED APPROACH

In Fig.1 a block diagram of main components of the proposed framework is shown. Each processing step is detailed in the following subsections.

2.1 Face Detection

First of all, facial images in the input video are detected and extracted. Face detection is a quite well handled task: in this paper, the well-known Viola-Jones object detector is exploited since it provides competitive face detection rates in real-time. Detected faces are then scaled to the largest one (to deal with scale changes) and radiometrically equalized in order to cope with different lighting conditions.

2.2 Eigenfaces

Face image data are then vectorized. At this point a new data representation is required in order to emphasize intra-person face similarity and to point-out inter-persons face differences. Moreover, considering that face data can have a high dimensionality, an efficient data reduction that preserves most of the amount of the embedded information is desirable. To this end, the Principal Component Analysis (PCA) is applied on face data as suggested in (Turk and Pentland,

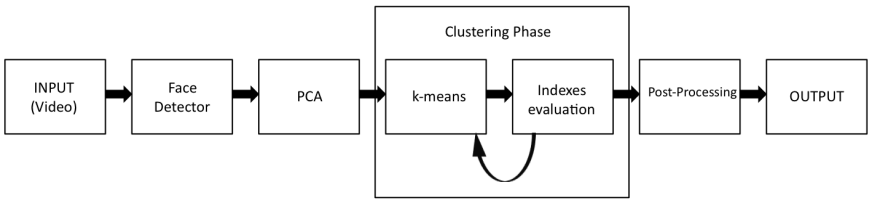

Figure 1: A block diagram of the proposed framework.

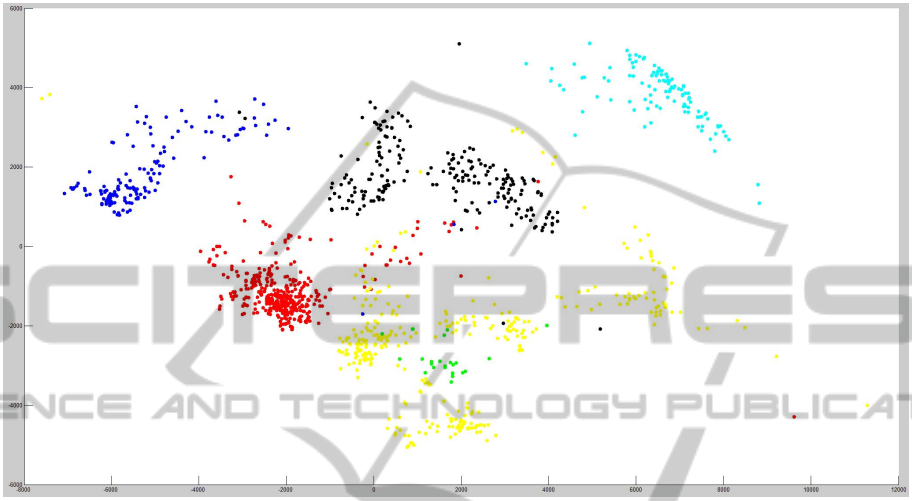

Figure 2: Plot of the scores related to the first two principal components in terms of Eigenface model.

1991). In this specific application context, PCA generates a set of eigenfaces, i.e. sets that represent the basis of a vectorial space where the input data are projected in order to achieve a better representation in terms of both inter-class and intra-class variation. The new representation is given by the weights associated to each eigenvector in order to get the complete initial set of data as a sum of (weighted) components. A score (the eigenvalue) is also associated to each eigenface and it indicates its importance in the representation of the initial data. This way, only the most relevant eigenfaces can be used for a more compact data representation. Fig.2 shows the representation of the first two components (i.e. the projection values of the initial data on the two most relevant eigenvectors) generated through the PCA on a set of face vectors corresponding to six different persons. Each color indicated a different person and it is evident that, even with only two eigenfaces, data shows a compact structure that is desirable for an efficient data clustering.

2.3 Face Clustering

From the eigenfaces theory comes that similar faces (i.e. faces belonging to the same person) have similar representation, i.e similar weights to the selected bases. Then a clustering algorithm on all of these weights can be efficiently performed and similar faces will belong to the same cluster.

K-means is a clustering technique that partitions a set of observation into k clusters so that each observation belongs to the cluster having the nearest mean. The only input this algorithm requires is the final number of clusters.

Unfortunately, in the considered application context, not having a prior knowledge implies that no supervisioned clustering is possible and, since the total number of people appearing in the video is unknown, a measure of k can not be a priori given. For this reason the well-known k-means algorithm is used in the following way: iteratively, k-means is run with a value of k increasing from 2 to a maximum (due only to a computational purpose, but it can be kept arbitrarily large), and then the best k is automatically selected.

In the most general case, this maximum value of k can be in a numerical interval ranging from 0 to N_{bound}, where N_{bound} is the number of frames in the video under consideration. It is straightforward to derive that this would bring to a huge number of iterations of the algorithm that could cause long delays in processing. To overcome this problem, the value of N_{bound} can be defined as a function that depends on the frame rate. That said, it is suggested to choose

N_{bound} as follows:

$$N_{bound} = \frac{T}{V.F.I.} \quad (1)$$

where $V.F.I.$ is the *Valid Frame Interval*, i.e. the minimum reasonable lapse of time in which a face should be present in order to be taken under consideration and T is the total video length (both measured in frames). In our test, $V.F.I.$ is defined as four time the video frame rate (i.e. a consistency of at least four seconds).

In particular the best k is chosen by evaluating several internal statistical indices, i.e. indices that are computed starting from the observation used to create clusters. Notice also that external indices can not be used since no a prior knowledge nor a prespecified data structure like a set of true known labels are available. In this paper, the investigated indices are the following: *Average Silhouette*, *Davies-Bouldin (DB)*, *Calinski-Harabasz (CH)*, *Krzanowski and Lai (KL)*, *Hartigan Index*, *weighted inter-intra (Wint) cluster ratio*, *Homogeneity-Separation*. For a more comprehensive treatment, refer to (Kaufman and Rousseeuw, 2009; Davies and Bouldin, 1979; Caliński and Harabasz, 1974; Krzanowski and Lai, 1988; Hartigan, 1975). The chosen criterion will be presented in section 3.

At the end of this step it is possible that the best number of clusters is not still defined since no satisfactory values are obtained during the whole iterative process. In that case, the hypothesis that only one person is present on the scene is made.

2.4 Post-processing

After clustering, each detected facial image is labeled as belonging to one of the k clusters found. Anyway some errors can occur: on the one hand the algorithm could create very small clusters, for example in correspondence of one or more false positive facial images detected by Viola-Jones algorithm. On the other hand, some segment could be split in case of miss-detection of the face detector. To overcome these problems and then to rightly determine the intervals of frames in which each person appears in the video a proper post-processing is introduced. It operates in a twofold manner (at a clustering level and, for a given cluster, at segment level) as follows:

1. a cluster is considered consistent if it classifies a person that is present in the scene for at least 4 seconds. All inconsistent clusters are removed;
2. two segments that have a temporal distance lower than 1.2 seconds are merged;
3. if a segment reveals a duration of less than 1.2 seconds but its neighbors are distant more than a frame number equal to 1.2 seconds, it is dropped from the segment list.

3 EXPERIMENTAL RESULTS

The proposed framework has been tested on several videos. The videos differ for number of people, people recurrences, lighting conditions, camera resolution, camera movements (quasi-static or continuously moving, like in the case of a mobile phone in the user's hand) and acquisition environments (indoor or outdoor). Each video has been, at first, processed by the face detector and then facial images are scaled, radiometrically equalized and finally projected, by the Principal Component Analysis, onto a feature space so that the element with greatest variance is projected onto the first axis, the second one onto the second axis and so on. At this point, for each video the minimum number of components to be retained for further processing has been set as the one able to preserve at least the 95% of the total variance of data. For example, for the fourth video, first 100 components overtake the threshold and are selected, like in Fig. 3. Reduced data are finally given as input to the generalized version of the k-means algorithm that, by the evaluation of a set of statistical indices, provides expected outcomes (i.e. the number of people and the intervals of frames in which each person appears).

In the first experimental phase the ability of the proposed framework to correctly detect the number of persons in the videos is tested. Table 1 reports the detailed results obtained for videos processed in this experimental phase. Each row lists a short description of the video (environment conditions i.e. indoor/outdoor, acquisition device i.e. mobile phone/camera, camera movements, i.e. M if the camera is in the hands of operator and then it continuously moves during recording, or S if the camera is quasi-static), the spatial resolution of the acquired images, the length of the video (in frames), the temporal resolution (fps), the total number of people appearing in the video and, in the last column, the number of people really detected by the proposed algorithm.

In the videos in rows 1-3 the proposed approach correctly detects the number of people that appear in it. In particular the first one is a video of size 1440×1080, with a frame rate of 30 fps and with 3600 frames. There are 8 persons, each one occurring once in the video. The video was acquired by a cam-

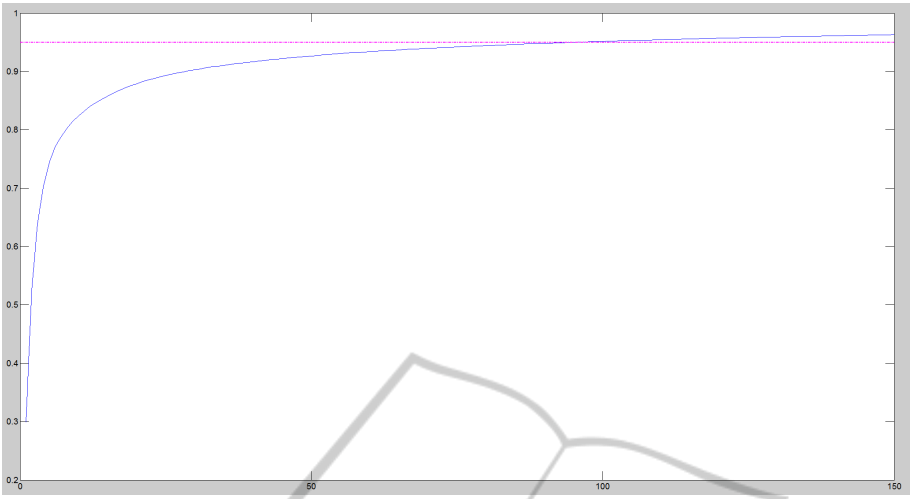

Figure 3: Number of components to be taken in order to preserve 95% of the total variance of data

era phone, in portrait configuration and in the hand of a walking person, so background totally changes, fast hand movements produced noise that is bigger due to the portrait configuration (the face detector will take a square region, that can include some black strip derived from pixels set black out of the border).

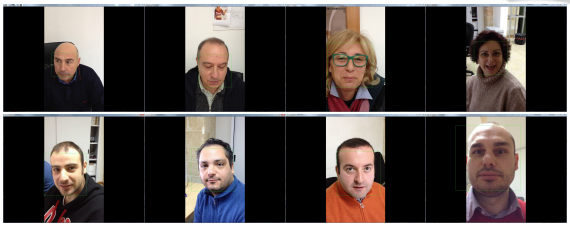

Figure 4: People present in the first test video.

Figure 5: Example of different poses for a person in the same video.

A snapshot for each person in the first video is reported in Fig.4. Fig.5 shows instead the great variability in pose, lighting condition, scale, background or blur that can occur among the facial images belonging to the same person. Finally, Fig.6 shows the values of the statistical indices for the first video. For figure clarity, only results until a value of N_{bound} equal to 20 are shown (using equation 1 it should be equal to 30).

From Fig.6 it is possible to perceive those computed indices that can bring to incoherent estimates of the best number of clusters. For this reason a good decision criterion should be defined. The criterion that best performs on the considered videos selects the value of k as the one that satisfies the Hartigan index selection criterion, i.e. to add a cluster while $H(k)>10$ and to estimate cluster number as the smallest $k \geq 1$ such that $H(k) \leq 10$, and at the same time has a corresponding Average Silhouette value greater than 0.45. For example, in the first video, the usage of the Hartigan index provides a wrong k value, but the corresponding Silhouette index reports a value that is lower than the threshold. The proposed decision criterion has avoided a wrong detection.

For videos reported in rows 1-3 of table 1 the detection of the right number of people fails. In fact, for the video in the fourth row, six persons are detected instead of four. In this case the system is not able to handle all the differences in pose, illumination and expression that strongly modify the face appearance of the facial images. In Fig.7 two persons that are erroneously split in four clusters instead of two are reported whereas in Fig.8 two persons with different appearances that are correctly classified into their respective cluster are shown.

In the second experimental phase the accuracy of the approach to determine the intervals of frames in which each detected person appears is tested. The accuracy of a segment is measured as the difference, in seconds, between ground truth start and end, compared with the computed ones. To this end, tables 2 and 3 report the accuracy of the segments extracted for the videos in the rows 3 (acquired in indoor) and 4 (acquired in outdoor) in table 1. The accuracy is very high for the indoor video (just two segments are slightly moved away from the corresponding ground truth data) and, as expected, decreases for the outdoor video where surrounding conditions are less constrained. Anyway the error, in most of cases, is below the second (often almost zero).

Figure 6: The investigated indices for the first test video.

Finally, in table 4, one example of how the post-processing works is reported. The considered video is the one reported in the second row in 1. In this video only one person is always present in the scene but he is not always detected by Viola-Jones algorithm. For this reason, before post-processing, more than one segment are created as reported (left part of table 4) and only post-processing application allows to get a unique segment that match the ground truth data (right part of table 4).

Summing up, the above tests show that the framework can effectively deal with pose, lighting condition, scale and blur variations. In most of the situations, the correct number of people was detected. Concerning the appearance interval for each person, the achieved accuracy is very high. The framework works better in indoor environments due to the extreme variation of images in outdoor. If these varia-

tions are high, a division of a cluster can sometimes happen.

Table 1: Experiments with 6 videos.

Video description	Resolution	Tot. Frames	FPS	Tot. People	Detected
Indoor, cellular, M	1080×1440	3600	30	8	8
Indoor, fixed camera, S	160×120	228	20	1	1
Indoor, cellular, M	1920×1080	4734	30	7	7
Outdoor, cellular, high reapp., M	640×352	3930	30	4	6
Indoor, cellular, high reapp., M	320×240	1913	29	5	6
Outdoor, cellular, high reapp., M	1080×1920	3219	29	4	5

Table 2: Indoor video.

Ground Truth		Detected		Error (sec)	
Start	End	Start	End	Start	End
4473	4610	4473	4610	0.00	0.00
3323	3611	3323	3611	0.00	0.00
3980	4070	3980	4070	0.00	0.00
66	429	88	392	0.73	1.23
4023	4136	4028	4136	0.17	0.00
4200	4418	4200	4418	0.00	0.00
3254	3419	3254	3419	0.00	0.00
916	977	916	977	0.00	0.00
1300	1396	1327	1396	0.90	0.00
1454	1576	1454	1576	0.00	0.00

Table 3: Outdoor video.

Ground Truth		Detected		Error (sec)	
Start	End	Start	End	Start	End
343	782	353	784	0.33	0.07
1089	1715	1128	1712	1.3	0.10
1805	2315	1892	2314	2.90	0.03
3682	3920	3682	3920	0.00	0.00
2465	3104	2471	3104	0.20	0.00

Figure 7: The two people divided into four clusters in the fourth test video.

Figure 8: Two people with different appearance but correctly classified into their respective clusters.

4 CONCLUSIONS

With this work a fully automated face indexing framework that works with home or camera phone videos was proposed. It automatically determines the number of people in the scene through different statistical indices and it also accurately reconstruct the intervals of frames in which each person appears. No initialization neither a priory knowledge about the scene contents are required. It has been experimentally proved that the proposed solution provides satisfactory results in different surrounding situations, even under a great variety of environments, poses and movements. Moreover, also low resolution videos, even taken from a mobile camera, can be successfully processed.

In order to further increase accuracy of the framework, several improvements can be made. For example, in order to cope with the the case in which a high variability in the facial pose originates more clusters on the same person, an head pose estimation technique and/or a face tracker based on appearance features could be added in order to lead the cluster generation.

Finally, our framework will be tested with publicly available databases.

Table 4: Unfiltered (left) vs. filtered (right) segments for one face.

Start	End
0	18
31	44
54	130
136	173
193	199
212	227

Start	End
0	227

REFERENCES

Arandjelovic, O. and Cipolla, R. (2006). Automatic cast listing in feature-length films with anisotropic manifold space. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 1513–1520. IEEE.

Bansal, N., Blum, A., and Chawla, S. (2004). Correlation clustering. *Machine Learning*, 56(1-3):89–113.

Caliński, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27.

Choi, J. Y., Plataniotis, K. N., and Ro, Y. M. (2010). Face annotation for online personal videos using color feature fusion based face recognition. In *Multimedia and Expo (ICME), 2010 IEEE International Conference on*, pages 1190–1195. IEEE.

Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (2):224–227.

Delezoide, B., Nouri, D., and Hamlaoui, S. (2011). On-line characters identification in movies. In *Content-Based Multimedia Indexing (CBMI), 2011 9th International Workshop on*, pages 169–174. IEEE.

Foucher, S. and Gagnon, L. (2007). Automatic detection and clustering of actor faces based on spectral clustering techniques. In *Computer and Robot Vision, 2007. CRV'07. Fourth Canadian Conference on*, pages 113–122. IEEE.

Görür, D. and Rasmussen, C. E. (2010). Dirichlet process gaussian mixture models: Choice of the base distribution. *Journal of Computer Science and Technology*, 25(4):653–664.

Hao, P. and Kamata, S.-i. (2011). Multi balanced trees for face retrieval from image database. In *Signal and Image Processing Applications (ICSIPA), 2011 IEEE International Conference on*, pages 484–489. IEEE.

Hartigan, J. A. (1975). *Clustering algorithms*. John Wiley & Sons, Inc.

Hu, W., Xie, N., Li, L., Zeng, X., and Maybank, S. (2011). A survey on visual content-based video indexing and retrieval. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 41(6):797–819.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254.

Kaufman, L. and Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*, volume 344. Wiley-Interscience.

Krzanowski, W. J. and Lai, Y. (1988). A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics*, pages 23–34.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, page 14. California, USA.

Pham, P., Moens, M.-F., and Tuytelaars, T. (2008). Linking names and faces: Seeing the problem in different ways. In *Proceedings of the 10th European conference on computer vision: workshop faces in'real-life'images: detection, alignment, and recognition*, pages 68–81.

Prinosil, J. (2011). Blind face indexing in video. In *Telecommunications and Signal Processing (TSP), 2011 34th International Conference on*, pages 575–578. IEEE.

Satoh, S., Nakamura, Y., and Kanade, T. (1999). Name-it: Naming and detecting faces in news videos. *MultiMedia, IEEE*, 6(1):22–35.

Sivic, J., Everingham, M., and Zisserman, A. (2009). who are you?-learning person specific classifiers from video. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1145–1152. IEEE.

Turk, M. and Pentland, A. (1991). Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86.

Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–511. IEEE.

Wold, S., Esbensen, K., and Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1):37–52.

Zhu, C., Wen, F., and Sun, J. (2011). A rank-order distance based clustering algorithm for face tagging. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 481–488. IEEE.