FACE RECOGNITION IN VIDEO USING DEMPSTER-SHAFER THEORY

S. Foucher, L. Gagnon

R&D Department, CRIM, 550 Sherbrooke West, Suite 100, Montreal (QC), CANADA, H3A 1B9
e-mail: {sfoucher, lgagnon}@crim.ca

ABSTRACT

We propose a new approach based on the Dempster-Shafer theory to increase identification performance of character faces in real films. We use a frame-based approach where faces are detected and identified independently in each frame without taking into account any temporal information. Statistical evidence about character faces is accumulated within the shot. Tests on documentary films show promising results both in terms of high detection and low false alarm rates.

1. INTRODUCTION

Character face identification in real videos like documentaries is a difficult task because of the high variability and complexity in visual content. Uncertainty and imprecision in statistical modeling arise naturally due to, for instance, missing learning data (i.e. unknown character faces) or variability in lighting condition, pose, etc. Such uncontrolled environment strongly limits the performance of the Bayesian approach which requires modeling and knowledge of the objects p.d.f.. Because of the content complexity, this knowledge is rarely reachable. Furthermore, total ignorance on the presence of an actor cannot be modeled within the Bayesian framework.

Dempster-Shafer theory enables us to manipulate both uncertainty and imprecision and has been successfully applied in data fusion problems (see [1] and references therein). We use it here to deal with the decision problem one gets once statistical evidence about faces have been gathered from a shot. Those evidences are cumulated from a *frame-based* approach where faces are detected and identified independently in each frame. We compare the identification results obtained using the Bayesian and Dempster-Shafer theory and evaluate them on a shot ranking application.

The paper is organized as follows. Section 2 briefly describes the underlying face detection and recognition algorithms that feed our Dempster-Shafer decision module. Section 3 reviews few concepts of the Dempster-Shafer theory. Section 4 presents the decision framework we use in this application. Tests and results are presented in Section 5.

2. FACE DETECTION AND RECOGNITION

We assume a video composed of K shots noted S_k. Frames in a shot are noted $F_t, t \in S_k$ with frame rate Δ_k. Each frame has H x W pixels of coordinates (x,y).

Face candidates are detected on each frame with the use a cascade of boosted classifiers implemented by Lienhart and Maydt [2]. On each face candidate V_n we assign a spatial likelihood

$$p(V_n \mid x, y, S_k, F_t) = \left(1 - e^{-\#\{hits\}/N_{\min}}\right) \times \Pi_{W_n}(x - x_n)\Pi_{H_n}(y - y_n) \quad (1)$$

H_n and W_n are the dimensions of the face bounding box. This likelihood depends on the number of hits given by the classifiers. In practice we use $N_{min} = 3$ so that false alarms have a low probability values.

Face recognition of a face candidate V_n is done with the HMM encoding procedure proposed by Nefian and Hayes [3]. In our application, the HMM classifier takes a decision among a set of hypotheses $\Omega = \{P_1, P_2, ..., P_L, fa\}$ where L is the number of persons and *fa* a false alarm class trained with non-face pictures. This identification process leads to a probability $p(P_l \mid V_n, S_k, F_t)$ for each of the trained face P_l.

3. DEMPSTER-SHAFER THEORY

A decision problem is composed of a set of mutually exclusive hypotheses called *frame of discernment* and noted Ω [4]. Information that contributes to the knowledge on the problem is captured by a *mass function*

$$m : 2^{\Omega} \rightarrow [0,1], m(\phi) = 0, \sum_{B \subseteq \Omega} m(B) = 1 \quad (2)$$

A particular mass $m(A)$ assigned to a subset $A \subseteq \Omega$ can be transferred to any hypothesis that makes up A, without knowing exactly which one. This degree of freedom reflects the degree of imprecision (or ignorance) on the

problem. Total ignorance is given by m(Ω)=1. Belief in a hypothesis *A* lies within an interval *[Bel(A),Pl(A)]* where

$$Bel: 2^{\Omega} \to [0,1], Bel(A) = \sum_{B \neq \phi; B \subseteq A} m(B) \quad (3)$$

$$Pl(A) = \sum_{B \cap A \neq \phi} m(B) = 1 - Bel(\overline{A}) \quad (4)$$

The interval *Bel(A)-Pl(A)*, called *belief interval*, can be interpreted as a degree of ignorance or imprecision.

3.1. Combination rule

The Dempster rule combines mass functions from two distinct information sources according to

$$m_{12}(A) = (m_1 \oplus m_2)(A) = \frac{\sum_{A_1 \cap A_2 = A} m_1(A_1) m_2(A_2)}{1 - C} \quad (5)$$

$$C = \tilde{m}_{12}(\phi) = \sum_{A_1 \cap A_2 = \phi} m_1(A_1) m_2(A_2) \quad (6)$$

K indicates the degree of conflict between the sources.

3.2. Discounting

When prior knowledge on the source reliability is known, belief function can be discounted [4]. The function

$$Bel_{\lambda}(A) = \begin{cases} \lambda Bel(A) & \text{if } A \neq \Omega \\ Bel(\Omega) & \text{if } A = \Omega \end{cases} \quad (7)$$

represents the discounted belief function, where λ is the degree of reliability. We use discounting here in order to introduce prior face detection probabilities.

3.3. Statistical mass functions

Mass functions can be derived from a statistical experiments [4]. We can sort the different decisions in statistical decreasing order $\Omega^{(0)} = \{d^{(0)}, d^{(1)}, ..., d^{(L+1)}\}$, where $d^{(0)} \in \Omega$ is the most probable hypothesis. For a *simple support belief structure* ($F = \{d^{(0)}, \Omega\}$) we have

$$m(A) = \begin{cases} 1 - p^{(1)} / p^{(0)} & \text{if } A = \{d^{(0)}\} \\ p^{(1)} / p^{(0)} & \text{if } A = \Omega \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

where $p^{(i)} \equiv p(d^{(i)})$. We use this belief structure to make a decision on the presence of an actor within a shot.

4. DECISION

4.1. Bayesian decision

Assuming decisions on the different frames are independent, we integrate over all the face candidates:

$$p(P_l \mid x, y, S_k, F_t) = \sum_n p(P_l \mid V_n, S_k, F_t) p(V_n \mid x, y, S_k, F_t) \quad (9)$$

We integrate these probabilities over all the sampled frames within the shot:

$$\begin{aligned} p(P_l \mid x, y, S_k) &= \sum_{t \in S_k} p(P_l \mid x, y, S_k, F_t) p(F_t \mid x, y, S_k) \\ &= \frac{1}{|S_k|} \sum_{t \in S_k} p(P_l \mid x, y, S_k, F_t) \end{aligned} \quad (10)$$

From this result we can derive two new probability distributions according to rows and columns:

$$\begin{aligned} p(P_l \mid x, S_k) &= \sum_{y \in [0, H-1]} p(P_l \mid x, y, S_k) p(y \mid x, S_k) \\ &= \frac{1}{H} \sum_{y \in [0, H-1]} p(P_l \mid x, y, S_k) \end{aligned} \quad (11)$$

A similar calculation leads to $p(P_l \mid y, S_k)$. From these two likelihood probabilities, we can derive an overall probability of presence of the actor in the current shot based on the geometrical mean:

$$p(P_l \mid S_k) = \sqrt{\left(\frac{1}{W} \sum_x p(P_l \mid x, S_k)\right)\left(\frac{1}{H} \sum_y p(P_l \mid y, S_k)\right)} \quad (12)$$

4.2. Belief theory with a simple consonant structure

According to Eq. (9), we have the following relation:

$$m(A \mid V_n, S_k, F_t) = \begin{cases} 1 - \dfrac{p(d^{(1)} \mid V_n, S_k, F_t)}{p(d^{(0)} \mid V_n, S_k, F_t)} & \text{if } A = \{d^{(0)}\} \\ \dfrac{p(d^{(1)} \mid V_n, S_k, F_t)}{p(d^{(0)} \mid V_n, S_k, F_t)} & \text{if } A = \Omega \\ 0 & \text{otherwise} \end{cases} \quad (13)$$

This mass function is discounted by the prior probability detection on V_n. From Eq. (7) we obtain:

$$m(A, V_n \mid x, y, S_k, F_t) = m(A \mid V_n, S_k, F_t) p(V_n \mid x, y, S_k, F_t) \quad (14)$$
$$, \forall A \subset \Omega$$

We consider each detected face candidate V_n as an independent source of information on the presence of a

particular hypothesis. The different sources are combined using Dempster rule:

$$m(P_l \mid x, S_k, F_t) = \bigoplus_n m(P_l, V_n \mid x, y, S_k, F_t) \tag{15}$$

For simple support belief structure, Eq. (15) becomes [1]

$$m(P_l \mid x, S_k, F_t) = \left(1 - \prod_{n \in \eta_l(x)} \left(1 - m(P_l, V_n \mid x, y, S_k, F_t)\right)\right) \times \left(\prod_{n \in \eta_l(x)} \left(1 - m(P_l, V_n \mid x, y, S_k, F_t)\right)\right) \tag{16}$$

where $\eta_l(x) = \left\{n \mid d^{(0)} = P_l \text{ et } p(V_n \mid x, y, S_k, F_t) > 0\right\}$ is the set of contributing face candidates that intersects along x. Similarly to Eq. (10) we integrate along the time axis:

$$m(P_l \mid x, S_k) = \frac{\Delta_k}{|S_k|} \sum_{t \in S_k} m(P_l \mid x, S_k, F_t) \tag{17}$$

5. EXPERIMENTS

5.1. Identification performance

We compare the Dempster-Shafer and Bayesian approaches on a complex shot containing two persons (noted P_2 and P_3 from left to right) side by side with a slow zoom and a fade-in/fade-out transition with a black and white picture (Fig. 1). These two persons are part of a characters set containing three other individuals ($L=5$).

Fig. 1. Key frames from a shot containing two persons (Marie-Louise and Pierre Saras, Le Fil Cassé, ©2001, NFB, All Rights Reserved)

From the row and column probability distributions we construct a 2-D likelihood function for the presence of character P_l at each location (x,y):

$$p(P_l \mid x, y, S_k) = \sqrt{p(P_l \mid x, S_k)\, p(P_l \mid y, S_k)} \tag{18}$$

From Eq. (18), we derive a spatial map of the person index l giving the maximum likelihood (Fig 2b and 3b):

$$\hat{l}(x, y) = \arg\max_l \left\{p(P_l \mid x, y, S_k)\right\} \tag{19}$$

Figure 2a shows the spatial probability distributions for the Bayesian case (Eq. (11)). No mode clearly emerge and the false alarm class dominates. On Fig. 2b, persons P_1 and P_5 (not present in the shot), have the maximum likelihood without any coherence in spatial domain. Using a simple support structure (Eq. (17)), we clearly observe two distinct probability modes for the two individuals P_2 and P_3 (Fig. 3a). In addition, the maximum likelihood gives a more accurate spatial domain estimation (Fig. 3b) for these two persons.

5.2. Shot ranking

In order to rank shots according to the presence of actor faces, we use the following posterior probability assuming equiprobable shots:

$$m(S_k \mid P_l) = \frac{m(P_l \mid S_k)}{\sum_k m(P_l \mid S_k)} \tag{20}$$

This can be interpreted as a degree of relevance of a shot to the actor face. On Table 1, we give the ranking of the first eleven shots (among *38*) for actors P_2 and P_3 (masses have been multiplied by *100*). Bold indicates shots where the actor is not present. Retrieval accuracy measured in terms of recall precision is *90.9%* for P_2 (resp. *63.6%* for P_3) and *54.5%* (resp. *18.3%*) for the Bayesian case.

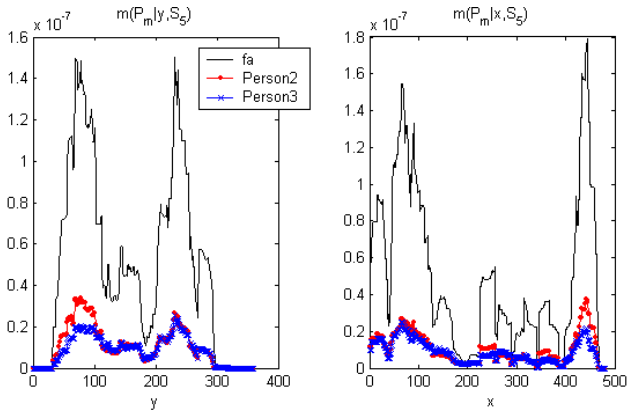
6. CONCLUSION

We use Dempster-Shafer theory to increase character face identification performance in real films. Compared to the Bayesian model, belief theory leads to better decisions and increased robustness against false alarms. The fact that belief theory uses non-singleton hypothesis and therefore allows a certain degree of imprecision in the decision process, produces more stable and reliable decisions. The current method has been used to encode face recognition in a system based on the MPEG-7 standard [5].

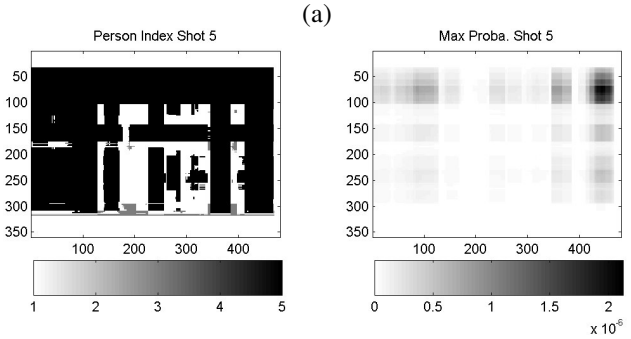
The same method can be used with other detection and recognition algorithms. In addition, joint use of detection and tracking algorithms could further increase the identification performance [6]. Other belief structures and decision rules can also be used [4].

ACKNOWLEDGEMENTS

This work is supported in part by CANARIE Inc. (www.canarie.ca), under the ARIM funding program. We thank the National Film Board of Canada for the videos.

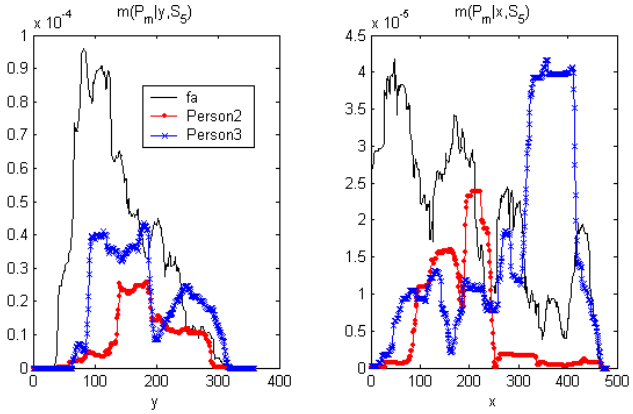


(a)

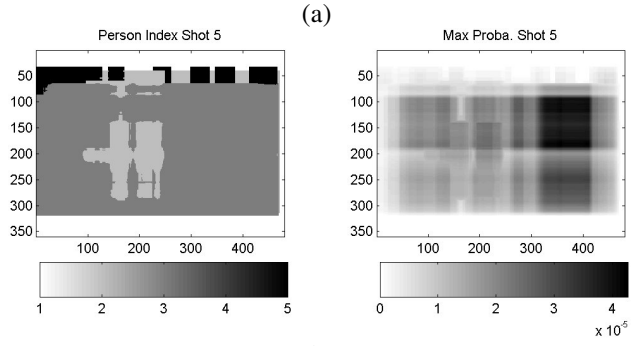


(b)

Fig. 2. Detection results with Bayesian approach.



(a)



(b)

Fig. 3. Detection results with Dempster-Shafer approach

k	Key frame	$m(S_k \mid P_2)$	k	Key frame	$m(S_k \mid P_3)$
24		11.97	*28*		20.84
22		9.95	*25*		19.28
12		9.71	*23*		6.69
19		9.46	*20*		5.87
26		7.98	***32***		5.76
25		7.14	*14*		4.90
6		6.54	***17***		3.20
8		6.22	*5*		2.73
27		3.45	*18*		2.62
13		3.30	***1***		2.33
29		2.66	***26***		2.29

Table 1. Top eleven most relevant shots for two different actors (Marie-Louise and Pierre Saras, Le Fil Cassé, ©2001, NFB, All Rights Reserved). Masse values have been multiplied by 100. Bold indicates shots where the actor is not present.

REFERENCES

[1] S. Foucher, J.-M. Boucher, G.B. Bénié, "Multiscale Classification and Filtering of SAR Images Using Dempster-Shafer Theory", *IGARSS'03*, Toulouse, 2003

[2] R. Lienhart, J. Maydt, "An Extended Set of Haar-like Features for Rapid Object Detection", *ICIP 2002*, Vol. 1, pp. 900-903, 2002

[3] A. V. Nefian, M. H. Hayes, "Face Recognition using an Embedded HMM", *IEEE Conference on Audio and Video-based Biometric Person Authentication*, pp. 19-24, 1999

[4] G. Shafer, *A Mathematical Theory of Evidence*, P. U. Press, Princeton, New Jersey, 1976

[5] L. Gagnon, S. Foucher, V. Gouaillier, C. Brun, J. Brousseau, G. Boulianne, F. Osterrath, C. Chapdelaine, J. Dutrisac, F. St-Onge, B. Champagne, X. Luc, "MPEG-7 audio-visual indexing test-bed for video retrieval", Proc. SPIE Internet Imaging 2004 (to appear)

[6] R.C. Verma, C. Schmid, K. Mikolajczyk, "Face Detection and Tracking in a Video by Propagating Detection Probabilities", IEEE-PAMI, Vol. 25, pp. 1215-1227, 2003