Linking Names and Faces: Seeing the Problem in Different Ways

Phi The Pham, Marie-Francine Moens, Tinne Tuytelaars

Linking Names and Faces: Seeing the Problem in Different Ways

Phi The Pham[1], Marie-Francine Moens[1], and Tinne Tuytelaars[2]

[1] Department of Computer Science, Katholieke Universiteit Leuven,
Celestijnenlaan 200A, B-3001 Leuven, Belgium
{PhiThe.Pham,Sien.Moens}@cs.kuleuven.be
[2] ESAT - PSI, Katholieke Universiteit Leuven,
Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
Tinne.Tuytelaars@esat.kuleuven.be

Abstract. In this paper we report on our experiments on linking names and faces as found in images and captions of online news websites. Whereas previously, the focus has been mostly on assigning names to the faces, we generalize this framework, exploiting the (a)symmetry between the visual and textual modalities. This leads to different schemes for assigning names to faces, assigning faces to names, and establishing name-face link pairs. On top of that, we investigate the use of textual and visual structural information to predict the presence of the corresponding entity in the other modality. This further improves the accuracy of the cross-media linking results.

1 Introduction

In this paper, we address the challenge of linking data across different modalities. In particular, we focus on the problem of linking names found in an image caption with the faces found in the corresponding image and vice versa. Such cross-media alignment brings a better understanding of the cross-media documents as it couples the different sources of information together and allows to resolve ambiguities that may arise from a single media document analysis (e.g. confusion between senior and junior George Bush). At the same time, it builds a cross-media model for each person in a fully unsupervised manner, which in turn allows to name the faces appearing in new images (with or without caption) or to visualize the people mentioned in new texts.

Because there are usually several names mentioned in the text and several faces shown in the image (see figure 1), and not all of the names have a corresponding face and vice versa, there are many possible alignments to choose from, making cross-media linking a non-trivial problem. However, analyzing a large corpus of cross-media stories (images with captions) the re-occurrence over and over again of particular face-name pairs provides evidence that they might indeed be linked to the same person. This is based on the assumption that the two modalities are correlated at least to some extent - a reasonable assumption for news stories where both modalities give a description of the same event.

This problem has been studied before (e.g. [1–3]). However, in earlier work the stress has always been on assigning names to the faces in the images. Here, our aim is to broaden this, exploiting the (a)symmetry between the two modalities. In a first model, we assume that the names of the text generate the faces in the image, in a second model we assume that the faces in the image generate the names in the text, and in a third model we consider the joint probability of the names and faces in order to compute the linking. We use here a standard Expectation Maximization algorithm. Additionally, because not all names of the text are equally important and the same is true for the faces in the image, the models can be corrected and possibly improved based on this salience information.

The remainder of this paper is organized as follows. We first describe some related work, followed by a description of the preprocessing steps. The next sections discuss the different linking models. Then, we shortly present the dataset used in the experiments and discuss some preliminary results.

2 Related work

Probably most similar to our work is the work of Berg et al. [1]. They were the first to study the problem of linking faces appearing in an image with the names mentioned in an associated text. They represent the faces in a face appearance space based on kernelPCA and cluster them using a Gaussian mixture model, where each component is assumed to correspond to a specific person. The parameters of the Gaussian mixture model are learnt together with the most probable assignments between detected faces and detected names based on an expectation maximization algorithm.

Our work is similar in spirit. However, we use a different face representation based on a 3D morphable model that has been fit to the image and as such allows to remove the effect of changes in pose and illumination. We avoid the restricting assumption that all faces of a specific person are Gaussian distributed. Instead, we vector quantize the feature space and use a piecewise linear approximation of the face distribution over these vector cells. Working with vector quantized features to represent the faces brings the additional advantage that now the textual (names) and visual (faces, or actually face clusters) representations are quite similar and both modalities can, to some extent, be interchanged. We do not impose a one-to-one relation between these names and face clusters though.

We generalize the framework of [1], playing with this intrinsic (a)symmetry between names and faces. Different models are proposed, depending on the task and focus, to assign names to faces (as done in [1]), to assign faces to names, or to identify pairs of names and faces.

Finally, we extend the framework by bringing in context information, which we refer to as *picturedness* (the probability that a person is actually in the corresponding picture, based purely on textual information) and *namedness* (the probability that a person is actually named in a text, based purely on visual information). Berg et al. also studied some simple text analysis cues, but not the other way around.

A related problem consists of finding faces of a single query person by exploiting both images and captions. A graph-based approach, where nodes correspond to faces and edges connect highly similar faces, has been studied by Ozkan et al. [4]. This method has later been refined by Guillaumin et al. [2], who also extend the method to deal with the multi-person naming problem. On their dataset, they outperform [1], which, they claim, is mostly due to the fact that, like us, they do not have the assumption of Gaussian distributions in feature space.

Finally, Jain et al. [3] proposed a scheme coined People-LDA. They incorporate context information from the text using a topic model inspired by Latent Dirichlet Allocation [5]. Topics are anchored on a specific person, using the faces as a guiding force. In fact, they do not use the person names at all, except for the initialization. They do not assume a Gaussian distribution for the faces, but instead use a generative model of the differences in appearance of two faces. This model is learnt beforehand, and the distribution over face space for a specific person is not updated after initialization. By including text information other than the person's name, they are able to assign the correct name even when it is not mentioned in the associated text, a limitation of all named entity based methods, including ours. However, in our experience, recognition of person names is relatively reliable - at least compared to the face detection/description steps, and it seems unwise not to use them.

3 Preprocessing

We focus on the text and the image that co-occur together. We call such an image-text pair a *story* x. The stories are part of a collection S.

3.1 Preprocessing of the texts

Detection of person names A first step is to recognize person names in the text. We use a named entity recognizer which is based on a maximum entropy classifier from the OpenNLP package[3], which we augmented with a gazetteer of names which were extracted from the Wikipedia[4] website.

Clustering of the person names In one text several mentions (e.g., "Al Gore", "former vice president", "he") might refer to the same person and form a coreference chain. Within one text this noun phrase coreference resolution follows the methods of the LingPipe[5] package. To group mentions of the same person across the stories, we use a dictionary of variant names in combination with a clustering of the coreference chains of a name, where the latter allows to resolve mentions of the single word "Bush" to "George W. Bush" and not to "Laura Bush". Then coreference chains of each text are clustered with a

[3] http://opennlp.sourceforge.net/
[4] http://en.wikipedia.org/
[5] http://www.alias-i.com/lingpipe/

hierarchical single link algorithm constrained by a threshold cosine similarity for cluster membership [6].

3.2 Preprocessing of the images

Detection and description of faces A parallel task regards the detection and description of the faces in the images. This is a challenging task under uncontrolled conditions, due to the wide variability in face appearance – especially because of changes in pose, illumination conditions, facial expressions, and partial occlusions. First, faces are detected using the OpenCV implementation of [7]. Next, we detect facial features [8] and use these as initial pose estimation for a 3D morphable face model [9] that is fitted to the data. Using such a 3D morphable model allows to estimate the pose and illumination parameters and to eliminate these irrelevant sources of variability. Also partial occlusions can be overcome this way. The model returns 40 person-specific texture components and 40 person-specific shape components, which together form the face descriptors used in this work. Unfortunately, these components are still affected by changes in facial expressions. As a result, assuming a Gaussian distribution in face descriptor space seems inappropriate (albeit not more inappropriate as for the face descriptors used in [1]).

Clustering of the faces To ease the later linking and to increase the symmetry between both modalities, we also cluster the faces, similar to the name clustering described above. Ideally, this would yield a single face cluster per person containing all the faces of that person and nothing else, so the problem could be reduced to linking a small number of names to a small number of face clusters. Unfortunately, in practice, after clustering several people are present within a single cluster, and the faces of a specific person are spread over several clusters.

The clustering of the faces groups similar faces based on the obtained face descriptors. Finding twice the face of the same person in an image is rare (although this might happen when a photograph or mirror image of the face is also present), so the clustering focuses on grouping faces across images and two faces of the same image are not allowed in the same cluster. We use a hierarchical agglomerative clustering algorithm and a cosine similarity metric, more specifically Group Average Clustering (GAC) with a predefined threshold of cluster membership similarity.

4 Overview of the different approaches

4.1 Assigning faces to names

One can think of the linking of names and faces as the problem of assigning suitable faces to names. For instance, given a text, the task could be to find a suitable illustration for it. In this case a text with names generates an image with faces. So, the task is to find a face f for a given name n. In each image-text pair

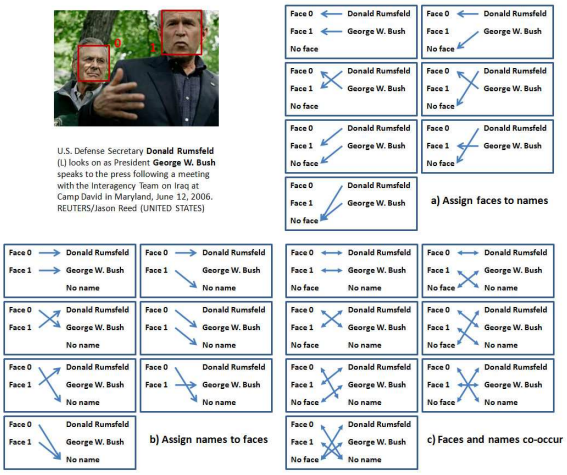

Fig. 1. An example of a story or image-text pair with its possible link schemes.

x_i, given N_i names, there are many possible link schemes a_j to assign F_i faces and a null face to these names (see figure 1), from which we have to choose the best one. The constraint for each link scheme is that a face must be assigned only to one name while the null face can be assigned to any name. When estimating the likelihood of a link scheme, the probability of a face given a name, $P(f|n)$, plays an important role.

4.2 Assigning names to faces

We can also inverse the above asymmetric assignment and assign names to faces. For instance, given an image, the task could be to describe the image content with text. In this case an image with faces generates a text with names. So, the task is to assign a name n to a face f. This is the usual way of looking at this problem, e.g. in the works of [1, 2]. In each image-text pair x_i, given F_i faces, there are again many possible link schemes a_j to assign N_i names and a null name to these faces (see figure 1). The constraint for each link scheme is that a name must be assigned only to one face, while the null name can be assigned to any face. When estimating the likelihood of a link scheme, the probability of a name given a face, $P(n|f)$, plays an important role.

4.3 Linking using the evidence of name-face co-occurrence

A stricter and more symmetric method is to use the joint probability, $P(f, n)$, instead of the conditional probabilities $P(f|n)$ or $P(n|f)$. $P(f, n)$ represents the probability that a certain name and a certain face co-occur. This can be obtained

by either using $P(n|f)$ or $P(f|n)$:

$$P(f,n) = P(f|n)P(n) = P(n|f)P(f) \tag{1}$$

It could be interpreted as follows. We no longer assume the names or faces to be given. Instead, both are drawn from a random distribution in one of the following ways. In a data set (e.g., todays news) a name occurs with a certain prior probability and given this name, we pick a face with a certain probability; or when a prior probability of the occurrence of a face is known, we pick a name to describe it. The prior probability and how it is estimated has an important influence on the result, compared to the likelihood function discussed in the previous sections. This could be considered as a Bayesian approach, where the two previous ones are more frequentist interpretations. In each image-text pair x_i, given F_i faces and N_i names, there are again many possible link schemes a_j to combine them. In this setting, a null name can be assigned to any face except for the null face and a null face can be assigned to any name except for the null name. An example of joint names and faces linking is shown in figure 1.

Discussion Obviously, the three link schemes described above are closely related. Each link configuration following one scheme can be transformed into an alignment following a different scheme by adding or removing assignments to the null name or null face. Nevertheless, it is important to clearly distinguish between the different settings, as it results in different initializations and normalizations during the optimization and also affects the reported results (measuring the accuracy relative to different entities).

5 Linking names and faces with an EM algorithm

For each story, we have to choose one link scheme among all possible schemes. One way to solve this names and faces matching problem is by implementing an expectation maximization (EM) algorithm. A hidden variable $\delta_{i,j}$ selects the most likely link scheme a_j for each story x_i. In practice, the $\delta_{i,j}$ are continuous variables allowing for soft decisions. C_i is the set of all possible links for image-text pair x_i. The EM iterates through two steps:

1. We estimate the likelihood of each link scheme a_j for each story x_i; and
2. We update the concerned probability distributions based on the estimated links.

We design several likelihood functions according to the different approaches distinguished above, as explained in the following sections.

5.1 Using $P(f|n)$ for the linking

We define the likelihood of the link scheme a_j for story x_i as follows:

$$L_{x_i,a_j}^{(n\to f)} = \prod_{\alpha} P(f_{\sigma(\alpha)}|n_\alpha) \tag{2}$$

where α is the index to the not null names in story x_i; $\sigma(\alpha)$ is the index to the faces (including the null face) assigned to these names. The complete log-likelihood of all stories S is:

$$\sum_{i \in S} \sum_{j \in C_i} \delta_{i,j} log(L^{(n \to f)}_{x_i, a_j}) \tag{3}$$

The expectation-maximization (EM) algorithm updates during the E-step $\delta_{i,j}$ as follows:

$$\delta_{i,j} = \frac{L^{(n \to f)}_{x_i, a_j}}{\sum_{l \in C_i} L^{(n \to f)}_{x_i, a_l}} \tag{4}$$

During the M-step the parameter $P(f|n)$ is maximized using soft counts:

$$P(f|n) = \frac{\sum_{i \in S} \sum_{j \in C_i} \delta_{i,j} m(a_j(n) = f)}{\sum_{i \in S} \sum_{j \in C_i} \delta_{i,j} m(n, a_j)} \tag{5}$$

where $m(a_j(n) = f)$ is 1, if face f is assigned to name n in the link scheme a_j, otherwise it is 0; $m(n, a_j)$ is 1, if the name n appears in a_j, otherwise it is 0.

5.2 Using $P(n|f)$ for the linking

To design the linking model using the information of how names are assigned to faces, we define the likelihood of the link scheme a_j for story x_i as follows:

$$L^{(f \to n)}_{x_i, a_j} = \prod_{\beta} P(n_{\sigma(\beta)} | f_{\beta}) \tag{6}$$

where β is the index to the not null faces in story x_i; $\sigma(\beta)$ is the index to the names (including the null name) assigned to these faces. The complete log-likelihood of all image-text pairs S is:

$$\sum_{i \in S} \sum_{j \in C_i} \delta_{i,j} log(L^{(f \to n)}_{x_i, a_j}) \tag{7}$$

The expectation-maximization (EM) algorithm updates during the E-step $\delta_{i,j}$ as follows:

$$\delta_{i,j} = \frac{L^{(f \to n)}_{x_i, a_j}}{\sum_{l \in C_i} L^{(f \to n)}_{x_i, a_l}} \tag{8}$$

During the M-step the parameter $P(n|f)$ is again maximized using soft counts:

$$P(n|f) = \frac{\sum_{i \in S} \sum_{j \in C_i} \delta_{i,j} m(a_j(f) = n)}{\sum_{i \in S} \sum_{j \in C_i} \delta_{i,j} m(f, a_j)} \tag{9}$$

where $m(a_j(f) = n)$ is 1, if name n is assigned to face f in the link scheme a_j, otherwise it is 0; $m(f, a_j)$ is 1, if the face f appears in a_j, otherwise it is 0.

5.3 Using $P(n, f)$ for the linking

In this approach, the information of how names and faces co-occur is used. The probability that the name n corresponds to the face f, i.e. $P(f, n)$ is defined using either the knowledge of how faces are assigned to names ($P(f|n)$) or how names are assigned to faces ($P(n|f)$) (See Equation 1). We define the likelihood of the link scheme a_j for story x_i as follows:

$$\begin{aligned} L_{x_i,a_j}^{(n,f)} &= \prod P(f,n) \\ &= \prod_{\alpha} (P(f_{\sigma(\alpha)}|n_\alpha)P(n_\alpha)) \\ &= \prod_{\beta} (P(n_{\sigma(\beta)}|f_\beta)P(f_\beta)) \end{aligned} \tag{10}$$

where α is the index to the not null names of story x_i; $\sigma(\alpha)$ is the index to the faces (including the null face) assigned to these names α; β is the index to the not null faces of story x_i; $\sigma(\beta)$ is the index to the names (including the null name) assigned to these faces β. The complete log-likelihood of all image-text pairs S is:

$$\sum_{i \in S} \sum_{j \in C_i} \delta_{i,j} log(L_{x_i,a_j}^{(n,f)}) \tag{11}$$

The expectation-maximization (EM) algorithm updates during the E-step $\delta_{i,j}$ as follows:

$$\delta_{i,j} = \frac{L_{x_i,a_j}^{(n,f)}}{\sum_{l \in C_i} L_{x_i,a_l}^{(n,f)}} \tag{12}$$

During the M-step: $P(f|n)$ or $P(n|f)$ used to estimate the likelihood $L^{(n,f)}$ is updated according to equations (5) and (9) respectively.

5.4 Use of a picturedness score

In the case of a text that generates certain images, the text might mention several distinct names, and, based on the structure of the text, not all of them have the same probability to occur in the image. We rank the person names of a story according to their probability of picturedness $P_{pictured}$ [10]. In this model we define the likelihood of the link scheme a_j for the image-text pair x_i as follows:

$$\begin{aligned} L_{x_i,a_j}^{(n^* \rightarrow f)} = & \prod_{\alpha, \sigma(\alpha) \neq NULL} (P(pictured_\alpha|t_{x_i})P(f_{\sigma(\alpha)}|n_\alpha)) \\ & \prod_{\alpha, \sigma(\alpha) = NULL} ((1 - P(pictured_\alpha|t_{x_i}))P(f_{\sigma(\alpha)}|n_\alpha)) \end{aligned} \tag{13}$$

where α is the index over names of story x_i; $P(pictured_\alpha|t_{x_i})$ is the probability that the name α appears in the image, given the text t_{x_i} of story x_i.

We compute $P(pictured_\alpha|t_{x_i})$ based on the salience and visualness of the name, where we assume that salient and visual names are more likely to occur in the image. We compute the salience score (ranging from 0 to 1) of a noun as a linear combination of the salience of the noun in the discourse and in the syntactic dependency tree of the sentence [11], where we learn the interpolation weights from a small training set. In [10] we computed a visualness score, which for person names equals one. $P(pictured_\alpha|t_{x_i})$ is the product of the salience and visualness of α, as we assume both scores to be independent, normalized by the sum of picturedness scores of all nouns in t_{x_i}. An extra regularization term of 0.01 is added to avoid scores equal to zero.

We hypothesise that the incorporation of picturedness factors into the likelihood function increases the accuracy of the name-face linking, especially for names and faces that rarely occur in the dataset, resulting in a relatively unreliable model for $P(f|n)$ or $P(n|f)$.

5.5 Use of a namedness score

Analogically, when an image generates a certain textual description, not all persons seen in the image are even likely to trigger a textual annotation. The largest faces in the image or the faces that appear in the center of the image are more likely to be described by names than other faces. Similarly, faces for which the 3D morphable model fitting procedure [9] converged to a stable solution are typically more reliable, bigger, sharper and more frontal, while an uncertain output of the fitting process often indicates errors in the face detection procedure or faces that are too small to yield accurate descriptors. As a result, also the confidence value returned by the fitting procedure can be used as a measure for namedness. Hence, we apply this confidence score (which we call "namedness") in our linking model in the hope that names are assigned more precisely to faces using this evidence. We modify the likelihood function of the link scheme a_j for a story x_i as follows:

$$L_{x_i,a_j}^{(f^* \rightarrow n)} = \prod_{\beta,\sigma(\beta)\neq NULL} (P(named_\beta|p_{x_i})P(n_{\sigma(\beta)}|f_\beta)) \prod_{\beta,\sigma(\beta)=NULL} ((1 - P(named_\beta|p_{x_i}))P(n_{\sigma(\beta)}|f_\beta)) \quad (14)$$

where β is the index over the faces of story x_i; $\sigma(\beta)$ is the index over the names assigned to the faces. $P(named_\beta|p_{x_i})$ is the confidence of the face β detected in the picture p_{x_i} of story x_i. It is also considered as the probability that a face is mentioned in the corresponding text.

5.6 Use of namedness and picturedness using the evidence of name-face co-occurrence

As the extension of the likelihood function $L^{(n,f)}$ (10), we incorporate the picturedness values of names or namedness value of faces with the wish to increase the accuracy of the names and faces association process.

We define the likelihood of the link scheme a_j for story x_i as follows (note that α and β range over not null names and not null faces respectively):

$$L^{(n^*,f^*)}_{x_i,a_j} = \prod_{\alpha,\sigma(\alpha)\neq NULL}(P(pictured_\alpha|t_{x_i})P(f_{\sigma(\alpha)}|n_\alpha)) \prod_{\alpha,\sigma(\alpha)=NULL}((1-P(pictured_\alpha|t_{x_i}))P(f_{\sigma(\alpha)}|n_\alpha)) \prod_\alpha P(n_\alpha) \quad (15)$$

$$L^{(n^*,f^*)}_{x_i,a_j} = \prod_{\beta,\sigma(\beta)\neq NULL}(P(named_\beta|p_{x_i})P(n_{\sigma(\beta)}|f_\beta)) \prod_{\beta,\sigma(\beta)=NULL}((1-P(named_\beta|p_{x_i}))P(n_{\sigma(\beta)}|f_\beta)) \prod_\beta P(f_\beta) \quad (16)$$

5.7 Initialization

To make EM converge to the true optimum, good initial values for the distributions $P(f|n)$, $P(n|f)$, $P(f)$ and $P(n)$ are essential. There are several initialization methods, of which we report one.

Initializing $P(f|n)$ based on initial clustering We have already identified in the stories clusters of similar names and clusters of similar faces that hopefully each represent one person name and one person face respectively, although in practice, they will be mixtures of different people, especially for the faces.

We adapt the method proposed by Mori et al. [12], where clustered segments of an image inherit all the words of the text that co-occur with the image. In our case the segments of the image are the detected faces, the words correspond to the person names extracted from the text. More specifically, we proceed as follows (see figure 2).

- Extract names and faces in every story x_i as explained in section 3;
- In each text, each name inherits all faces from the corresponding image;
- Make name clusters from all texts as explained in section 3.1;
- The faces that are inherited by one name cluster are grouped when they belong to one face cluster obtained in 3.2;
- The $P(f|n)$ distributions are estimated by counting the relative frequencies of the face clusters inherited by a given name.

Initializing $P(n|f)$ based on initial clustering Estimating $P(n|f)$ is similar to estimating $P(f|n)$, except that references to names and faces and to their clusters are switched in the above algorithm.

There is still the problem of estimating $P(f|n)$ if f is the null face, and of $P(n|f)$ if n is the null name, for which the name and face clusters obtained by considering all the stories give no information. We estimate these by considering individual stories. $P(f|n)$ is then computed as $1/(F_i + 1)$; $P(n|f)$ as $1/(N_i + 1)$, where F_i is the number of faces in story x_i, N_i is the number of names in x_i, and 1 refers to the null face or null name respectively. Alternative initialization schemes are currently being studied.

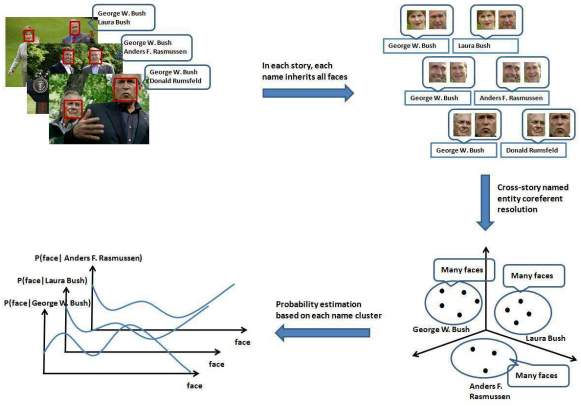

Fig. 2. Concept of the initialization method for estimating $P(f|n)$.

Initializing $P(n)$ and $P(f)$ Given the name clusters obtained in section 3.1, $P(n)$ is computed by maximum likelihood estimation of the occurrence of a name among all names in the collection. Given the face clusters obtained in section 3.2, $P(f)$ is computed by maximum likelihood estimation of the occurrence of a face among all faces in the collection.

5.8 Computing the most likely link scheme

The EM algorithm is iterated until the value of the complete log likelihood function for all stories S no longer increases. Then, for each story x_i, the link scheme a_j among the possible link schemes is chosen for which the corresponding $\delta_{i,j}$ is maximum.

6 Test set and results

We evaluate on image-text pairs selected from a set of captioned news photographs collected from the Yahoo! News website[6] of which the names and the faces have been manually linked. This dataset consists of 11820 stories, each contains on average 2.00 person names and 1.12 faces with a standard deviation of 0.0097 and 0.0032 respectively. After preprocessing there are 13233 faces clustered into 749 face clusters. The obtained face clusters are rather noisy. With the threshold (set empirically based on a small validation set) of face similarity measured with a cosine metric used for the reported results, the normalized mutual information (NMI) between the obtained face clusters and the ground truth face clusters is only 33%. NMI measures the mutual information between

[6] We use the "Yahoo! News" dataset of [1], with ground truth annotations from the "Labeled Faces in the Wild" dataset thanks to [13].

Accuracy	After 1 iteration	full EM
Likelihood type $L^{(n\to f)}$	71.15%	71.24%
Likelihood type $L^{(f\to n)}$	58.71%	60.15%
Likelihood type $L^{(n,f)}$ using $P(f\|n)$	69.51%	69.58%
Likelihood type $L^{(n,f)}$ using $P(n\|f)$	60.98%	62.70%
Likelihood type $L^{(n*\to f)}$	71.49%	71.61%
Likelihood type $L^{(f\to n)}$	57.83%	59.16%
Likelihood type $L^{(n*,f*)}$ using $P(f\|n)$	69.58%	69.70%
Likelihood type $L^{(n*,f*)}$ using $P(n\|f)$	60.49%	61.90%

Table 1. Accuracy of the name-face linking of the "Labeled Faces in the Wild" dataset.

two cluster sets [14]. Optimizing this threshold in a separate experiment only yielded a maximum of 48% NMI score. There are 23733 names clustered into 9793 name clusters. The person name recognition has an accuracy of 81% on a validation set of 100 stories.

We measure the accuracy of the faces assigned to names ($L^{(n\to f)}$, $L^{(n*\to f)}$) as the number of correct assignments of faces to names divided by the number of names, and the accuracy of the names assigned to faces ($L^{(f\to n)}$, $L^{(f*\to n)}$) as the number of correct assignments of names to faces divided by the number of faces. In case of link schemes based on evidence of name-face co-occurrence ($L^{(n,f)}$ using $P(f|n)$, $L^{(n,f)}$ using $P(n|f)$, $L^{(n*,f*)}$ using $P(f|n)$, $L^{(n*,f*)}$ using $P(n|f)$), the accuracy is the number of correct face-name pairs over all face-name pairs. Correct or incorrect assignments of the null face or null name influence the scores. In an alternative evaluation we exclude links with the null face or null name in the accuracy computations.

The results of the name-face linking in the stories in terms of accuracy after the first iteration and after convergence of the EM algorithm are presented in table 1. For these results, we assume that in our dataset there is a minimum correlation between names of the texts and faces of the image, i.e., not all faces are assigned to null name and not all names are assigned to null face. Consequently for all link models, we reduce the ambiguitiy by ignoring this special assignment scheme where all names are assigned to null face and all faces are assigned to null name. We still use the accuracy measures as described in the previous paragraph as this heuristic does not affect the number of names or faces that have to be assigned respectively a face or name. However, for datasets that lack this correlation, our heuristic could harm the accuracy results.

It is clear that the initial linking (after just one iteration) already yields very good results, which are only slightly improved by the application of the EM algorithm. This is due to the fact that the initial result already incorporates a linking based on the complete dataset.

It is easier to correctly assign faces to names ($L^{(n\to f)}$ and $L^{(n,f)}$ using $P(f|n)$) than vice versa ($L^{(f\to n)}$ and $L^{(n,f)}$ using $P(n|f)$). This might be explained by the data set and the proposed initialization method. In the stories there are usually more names than faces, so a name initially inherits few faces, while a

face might inherit more names, making the initial linking more ambiguous for the latter case.

Note that these results also include the assignments to null name and null face. When only evaluating the linking between real names and faces, the performance slightly decreases for the model based on $L^{(n \to f)}$ and slightly increases for the model based on $L^{(f \to n)}$ (in the order of $1-2\%$). This is explained by the acccuracy computations. For instance, if you have many names but few faces, there is a good chance of correctly assigning the null face to the names that have no real face, an advantage, which disappears when only considering real names and real faces. When only evaluating the linking between real (not-null) names and faces, the different evaluation schemes become equivalent, and so we can conclude that systematically better results are obtained when using $P(f|n)$ instead of $P(n|f)$. This also becomes evident when comparing the two different schemes for computing $L^{(n,f)}$ and $L^{(n*,f*)}$.

From the results the positive influence of the picturedness value is present, but is rather modest This might be due to the performance of the person name recognition which was lower than expected in this dataset. The detection of visual cues that signal a possible textual description might be improved. We are currently performing tests based on centrality and location of a face in the image.

Berg et al. [1] selected 1000 faces randomly from the "Faces in the wild dataset" and evaluatede the correctness of the linking with names in the text. They report a 56% accuracy when no language information from the text was used, and 72% accuracy when in each step of the EM iteration a better language model was learned. We learn a language model (socalled "picturedness") once prior to the alignment computations. Note, that these authors do not evaluate all links in one story. Instead, they randomly select faces in the dataset and evaluate the accuracy of the names assigned. We performed a similar experiment and obtained an accuracy of 70%. In their language model Berg et al. [1] use textual cue phrases such as "left" or "right" in the context of a name that signal the presence of a face, which we have ignored in our models so far.

7 Conclusions and future work

We reported here on experiments in name-face links on real-world captioned news photographs. Our preliminary results indicate that correctly detecting the links is not a trivial task. In particular, we have studied the effect of the (a)symmetry between the two modalities, text and image. Our methods achieve a 71% accuracy where we evaluate the links between all names and faces in the "Labeled Faces in the Wild" dataset.

Many other link or alignment algorithms - possibly inspired by methods used in machine translation for aligning words and phrases in bilingual corpora - can be implemented and tested. In addition, linking of faces and names in video streams might be investigated. Current interest in content recognition in text and

images promotes cross-media alignment of detected objects, actions, complete scenes, scenarios or concepts, offering many challenging research questions.

Acknowledgements

The authors acknowledge the support of European Project CLASS (EU IST FP6-027978), the Flemish SBO project AMASS++ (IWT-SBO-060051), and the Fund for Scientific Research Flanders.

References

1. Berg, T., Berg, A., Edwards, J., Maire, M., White, R., Teh, Y., Learned-Miller, E., Forsyth, D.: Names and faces in the news. In: CVPR. Volume 2. (2004) 848–854
2. Guillaumin, M., Mensink, T., Verbeek, J., Schmid, C.: Automatic face naming with caption-based supervision. In: CVPR. (2008)
3. Jain, V., Learned-Miller, E., McCallum, A.: People-lda: Anchoring topics to people using face recognition. In: Proceedings International conference on computer vision. (2007)
4. D. Ozkan, P.D.: A graph based approach for naming faces in news photos. In: CVPR. (2006)
5. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research **3** (2003)
6. Pham, P.T., Moens, M.F., Tuytelaars, T.: Cross-media alignment of names and faces. Technical report, Katholieke Universiteit Leuven (2008)
7. Viola, P., Jones, M.: Robust realtime object detection vector quantization. International Journal of Computer Vision **57** (2004) 137–154
8. Everingham, M., Sivic, J., Zisserman, A.: 'hello! my name is... buffy' - automatic naming of characters in tv video. In: Proceedings of the 17th British Machine Vision Conference. (2006) 889–908
9. M. Desmet, R. Fransens, L.V.G.: A generalized em approach for 3d model based face recognition under occlusions. In: CVPR. Volume 2. (2006) 1423–1430
10. Deschacht, K., Moens, M.F.: Text analysis for automatic image annotation. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. (2007) 1000–1007
11. Charniak, E.: A maximum entropy-inspired parser. In: Proceedings of the First Conference of the North American Chapter of the Association for Computational Linguistics. (2000)
12. Mori, Y., Takahashi, H., Oka, R.: Automatic words assignment to images based on image division and vector quantization. In: Proceedings of RIAO' 2000 Content-Based Multimedia Information Access. (2000)
13. Huang, G.B., Rameh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst (2007)
14. Strehl, A., Ghosh, J., Cardie, C.: Cluster ensembles - a knowledge reuse framework for combining multiple partitions. Journal of Machine Learning Research **3** (2002) 583–617
15. Baldonado, M., Chang, C.C., Gravano, L., Paepcke, A.: The stanford digital library metadata architecture. Int. J. Digit. Libr. **1** (1997) 108–121