Exploiting visual context to identify people in TV programs

Thomas PETIT[1,2], Pierre LETESSIER[1], Stefan DUFFNER[2], and Christophe GARCIA[2]

[1] Institut National de l'Audiovisuel, Bry-sur-Marne, France
{tpetit, pletessier}@ina.fr
[2] Univ Lyon, INSA Lyon, LIRIS (UMR 5202 CNRS)
{thomas.petit, stefan.duffner, christophe.garcia}@liris.cnrs.fr

Abstract. Television is a medium that is implicitly highly codified. Every TV program has its own visual identity that is often rich in information; most of the time, a single frame extracted from a TV broadcast contains enough information for a human agent to determine the genre of the program, and sometimes even to predict who is likely to appear in it. Our goal is to exploit the visual context of TV programs to help identify the people appearing in them.
In this work, we introduce a new dataset of over 10M frames extracted mainly from french TV programs and aired between 2010 and 2020. We also present an original approach for deep similarity metric learning in order to learn a descriptor that effectively captures the visual context of a TV program and helps to recognize the subjects appearing in the program.

Keywords: dataset · similarity measure, · visual context · television

1 Introduction

Automatic facial recognition efficiency has increased considerably in the last decade and can now achieve impressive results. However, these state-of-the-art models may still make some mistakes on faces that are very hard to distinguish. In most cases, humans would not make the same mistakes as they would have access to much more information.

Among the additional information that the human brain uses to identify people, the visual context, i.e. all the visual information except for the face to identify, is particularly useful. We know, for example, that a human agent is able to achieve an accuracy score of 94.27% on the LFW protocol [3], which is a face verification protocol, when all faces have been masked [5], meaning using only the surrounding visual context.

We also know that the television is a very codified medium; given only a few static frames, a human is most of the time able to say from which kind of program they have been sampled, whether it is a sport match, a newscast, a

political debate, and so on. We believe that this knowledge carries much information about the people possibly appearing in that program and could be used to disambiguate the cases in which facial recognition systems fail. Our goal is to exploit the visual context to improve face retrieval and face verification in a dataset of TV programs.

In this paper, we present two contributions: 1) we introduce a new large-scale dataset of over 10M images from TV programs aired between 2010 and 2020; 2) we propose a deep metric learning approach for visual context building an effective descriptor specific to TV. We experimentally show that this new descriptor can be used jointly with state-of-the-art face descriptors to improve the performances over a face verification task and a face classification task when such a visual context is available.

2 Related Work

Exploiting contextual information in order to improve facial recognition is an approach that has already been studied. However, as the available contextual information can be very specific to a given problem, the different approaches and solutions can differ a lot.

Some studies focused on exploiting a social context by identifying the relationships between the subjects or the events they appear in. This approach is particularly suitable for social networks where the relationships between the users are explicit [11] but have also been applied to other kinds of datasets such as movies [4] or TV shows [6].

Other papers chose not to limit the visual appearance of the subjects to their faces but also took advantage of their clothes by extracting features from different body parts and merging them together. This approach, which is particularly suited for person re-identification in a unique event, has been applied on photo albums [15] and on movies [4].

A different problem which is worth mentioning here because of its proximity with our approach is the genre classification for TV videos. The goal is to classify TV videos into a few different genres. Some models rely on automatically learned features, like Varghese et al. [12] classifying videos into different genres ("news", "entertainment" and "sport") using features learned on the SUN dataset [14]. Other models, on the other hand, use handcrafted features like Daudpota et al. [1] who rely on the number of scenes and quantify the motion of the subjects in order to classify videos as "talk shows" or "others".

Several datasets have already been proposed to deal with the scene categorization problem in a general setting. Two of them in particular are widely used: the SUN [14] and the Places365 dataset [17]. They contain a large number of indoor and outdoor scene categories. These scene categories are very diverse; Places365, for example, includes categories like "igloo", "synagogue" or even "stables". This large number of various categories is very interesting but does not make these datasets particularly suited to learn a visual descriptor of TV frames, as most categories are not related to TV programs.

3 Dataset

3.1 Motivation

Our goal is to be able to exploit the visual information in a video frame other than the facial appearances to help identify the people appearing on TV programs. We want to be able to extract continuous feature descriptors from video frames that carry information about the context of the program, and hence about the peoples appearing in those frames.
Existing datasets oriented towards context recognition have been designed in order to classify the images. Some datasets of locations images like SUN [14] or Places365 [17] are quite exhaustive in terms of labels, however they do not allow to capture the semantic proximity that can exist between two different classes, like between an airfield and an airport terminal, or between the ocean and a harbour. Also their distribution is quite different from what is to be expected in a dataset of TV frames. Some other datasets which are more TV specific only focus on classifying TV programs in a few classes like news, sports, music, and so on, but they fail at capturing the diversity within each of these classes and the semantic relationships that can exist between them.

3.2 Dataset structure

We introduce a dataset of 10,684,217 frames of TV programs aired between 2010 and 2020, mostly on the French TV, but not only. It is public and is completely available[3]. This dataset covers the diversity of visual contexts prone to appear on television with frames selected from a large number of TV programs of all sorts, such as news, sports, entertainment, talk shows, and so on. For legal reasons as well as practical reasons that will be detailed below, all the faces have been blurred in this dataset, so as to be unrecognizable. Fig. 1 shows a few examples of frames from our dataset.

This dataset is unlabeled. However, it comes with a list of 4,362,818 pairs of frames where at least one individual appears on both frames.

4 Visual context metric learning

The dataset has been built in order to be able to compute a continuous feature descriptor from a static image that captures its visual context and that can help identify the people depicted in it. To this end, we propose a deep metric learning approach and organize our dataset in triplets including one anchor, one positive element and one negative element, so that a similarity metric can be learned using the triplet loss or the TSML loss. This approach has been proved to be efficient for continuous visual descriptor learning [7], and we adapted it to specifically learn visual context as explained in the following.

[3] URL to the dataset will be released after the review process

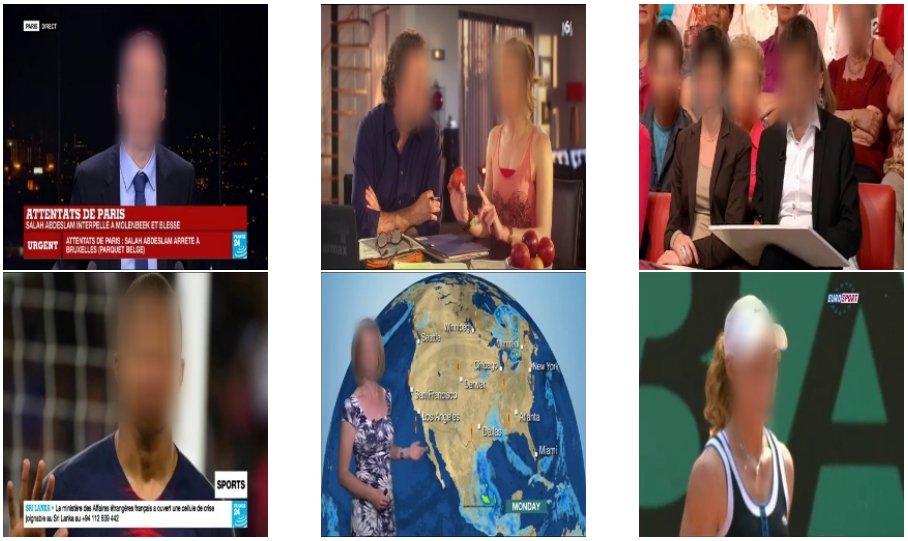

Fig. 1. Sample of the frames from our dataset. The visual contexts are various and reflect the diversity of the programs one can see on TV: news, entertainment, sport, weather forecast, and so on.

4.1 Triplet formation

The difficulty lies in being able to define what makes a positive or a negative pair of frames. The purpose of this dataset is to help identify the persons appearing in the frames. Thus, we decided to rely on person identities to build our positive and negative pairs of frames, needed to form the (anchor, positive, negative) triplets.

Positive pairs We consider one pair of frames as positive context-wise if we are able to identify at least one person that appears in both frames. Given the impossibility to label manually all of the faces appearing in over 10M frames, we performed this automatically. We first detected all of the faces in the original frames of our dataset (not blurred) and computed the corresponding facial features descriptors. We then formed our pairs using a selective distance threshold between the faces to assert they do in fact belong to the one same person. Some examples are displayed on Fig. 2. The facial features model we used is a ResNet18 architecture trained on the VGGFace2 dataset. It achieves a 98.98% score on the LFW protocol [3].

Negative pairs A common and effective strategy in similarity metric learning setting is to focus on hard negative examples during training. For example, for facial features learning, comparing similar identities help differentiating them [10, 13, 9, 7]. This often implies selecting negative pairs with similar embeddings.

Fig. 2. Examples of two positive pairs from our dataset.

Our problem, however, differs. If we can consider a pair as being positive when a common person appears in both frames, the absence of such a person is not enough to consider a pair as being negative. We might for example, sample two frames from the same program where no common people appears on both frames. We do not expect this pair to be considered as negative since the context is not expected to be any different within a unique program from one scene to another. The same can apply to two different programs with a very similar context but no common participant.

This make such an adversarial sampling very difficult to apply in our case as it could lead to sampling too many false negative pairs. For this reason, we decided to sample the negative elements of our triplets randomly and to rely on the large size and on the diversity of our dataset to make false negative pairs highly unlikely.

4.2 Model learning

Architecture To train our visual context model, we use a Resnet50 architecture[2] pre-trained on Places365 [17][4]. The last classification layer, with 365 nodes, is replaced with a 16-dimensional layer. The input of this network are 256×256 images.

Loss function The pretrained model is fine-tuned using the Triangular Similarity Metric Learning (TSML) loss [16]. This loss function is similar to the widely used triplet loss introduced in [7]. However, we observed that the performance using the TSML loss is slightly better. Hence, in this paper, we report only the TSML results.

About the face blurring We mentioned earlier that all faces in our dataset have been blurred, in particular for legal reasons. It appears it is also a practical choice. A first model has been trained similarly using the same dataset on which faces had not been blurred. Its performances were satisfying; however, we noticed using some visualization techniques like the Smoothgrad algorithm [8] that this model focuses primarily on the faces appearing on the images and not on the

[4] https://github.com/CSAILVision/places365

background as was desired (see Fig. 3). Moreover, its performances decreases when applied on frames where faces were blurred, which proves that it learned to recognize the faces as well as to recognize the visual context. This is not surprising given our triplet formation strategies, described above, where at least one person appears in both images of a positive pair. Blurring the faces in the dataset helped to largely avoid this issue.

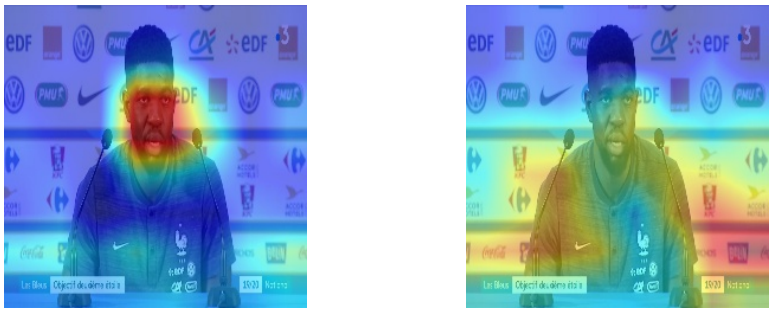

Fig. 3. When using a variant of our dataset where faces are not blurred, the trained model focuses mainly on the actual faces and not on the surrounding context (left). This is not the case when trained on our dataset with blurred faces (right).

5 Experimental evaluation

We split our dataset in three subsets:

- a training set, containing 4,357,969 positive pairs and 4,331,132 more elements to form the negative pairs of the triplets during training
- a validation set, with 2,456 triplets
- and a test set containing 2,393 triplets

After training our model on our training set, and using the validation set for early stopping, we evaluated it on our test set and compared it to other existing models.

5.1 Visual context metric evaluation

To evaluate our model, we split our test set into 5 and use it to perform a 5-fold cross-validation to determine the best threshold to classify a pair as positive or negative, i.e. same or different context. The overall accuracy displayed in Table 1 is the average accuracy obtained over the 5 folds $\pm$ the standard deviation.

We compare our model with the pre-trained Places365 models. For these pre-trained models, we use either the 365-dimensional outputs of the classification layer or the 2048-dimensional output of the previous layer (which is the input of the classification layer itself).

Table 1. Average accuracy ± standard deviation with 5-fold cross-validation over our test set, for our model and for pre-trained Places365 models

Model	Ours	Pl365 Resnet50	Pl365 Resnet50 2048-d layer	Pl365 Densenet161
Acc.	**85.17 ± 1.46%**	75.34 ± 0.68%	76.72 ± 0.65%	74.70 ± 0.68%

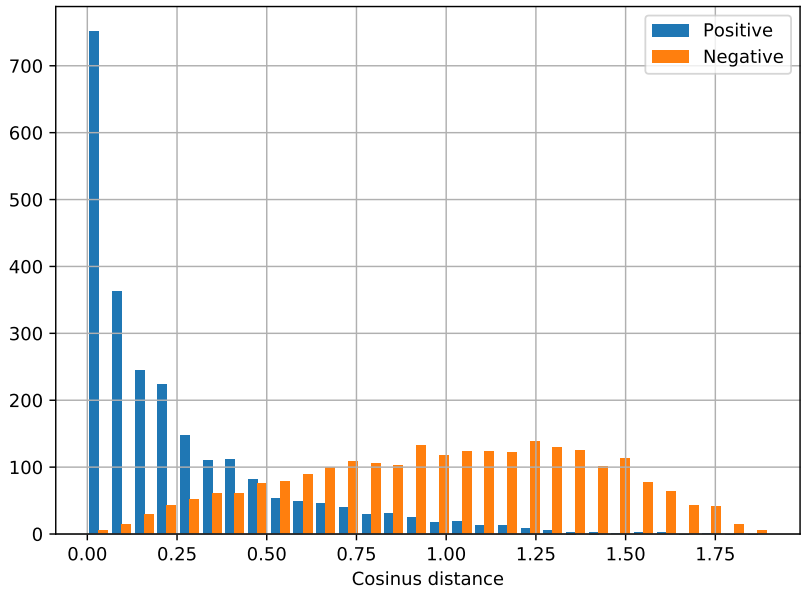

Fig. 4. Distributions of the distances of the positive and negative pairs of the test set computed with our visual context model.

Figure 4 shows the distribution of the cosine distances of the positive and negative pairs of the test set using our model. We can observe that the distribution of the negative pairs matches what would be expected from a random distribution uniformly distributed on the hyper-space. The distribution of the positive pairs, however, is much more concentrated towards low distance values.

5.2 Evaluation on a face verification task of doppelgangers

We evaluate the ability of our model to help recognize people when used jointly with facial features descriptors. In order to do this, we sample several face pairs to form a face verification task. The positive pairs are sampled in the same way as for the training set, i.e. with two face images of the same person. The negative pairs, however, have not been sampled randomly; we selected hard examples where both members of a pair are visually similar but are not the same person (see Fig 5). All images have been sampled from TV programs that do not appear

in the training, validation or test set introduced above. These face verification images are available with our dataset.

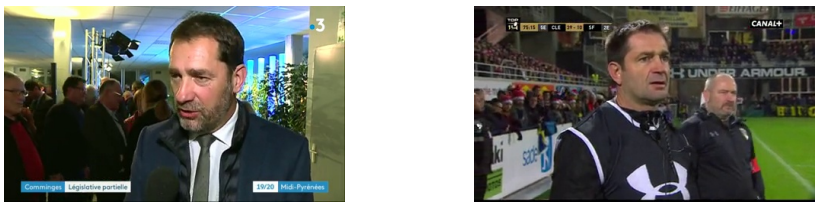

Fig. 5. Example of doppelgangers pair used for the face verification task.

We use a total of 2922 pairs distributed into 2 splits. The first split is used to learn the best weight to combine the facial descriptors distances and the contextual descriptors distances and to learn the best threshold to classify pairs as positive or negative. These parameters are applied on the second split to get an accuracy score. The two splits are then swapped and this operation is repeated.

In Table 2 are displayed the average accuracy scores over the two splits. We observe that the combination of both facial and visual context descriptors achieves a better performance than the facial descriptors alone.

Table 2. Average accuracy over both splits of the doppelgangers verification task

Input	Faces only	Context only	Faces + context
Accuracy	85.87 ± 0.03 %	65.86 ± 0.47%	**87.10 ± 0.17%**

6 Qualitative results

In order to illustrate how our model performs, Fig. 6 displays a few sample images and their nearest neighbors in our test set.

From these examples, we can see that frames from newscast, cartoons or weather forecasts are close to each other, respectively, even when coming from different TV channels.

7 Conclusion

In this paper, we presented a new approach of leveraging visual context information in TV programs for person recognizing. We introduced a new dataset of over 10M frames from TV programs, that have been broadcast over an entire decade.

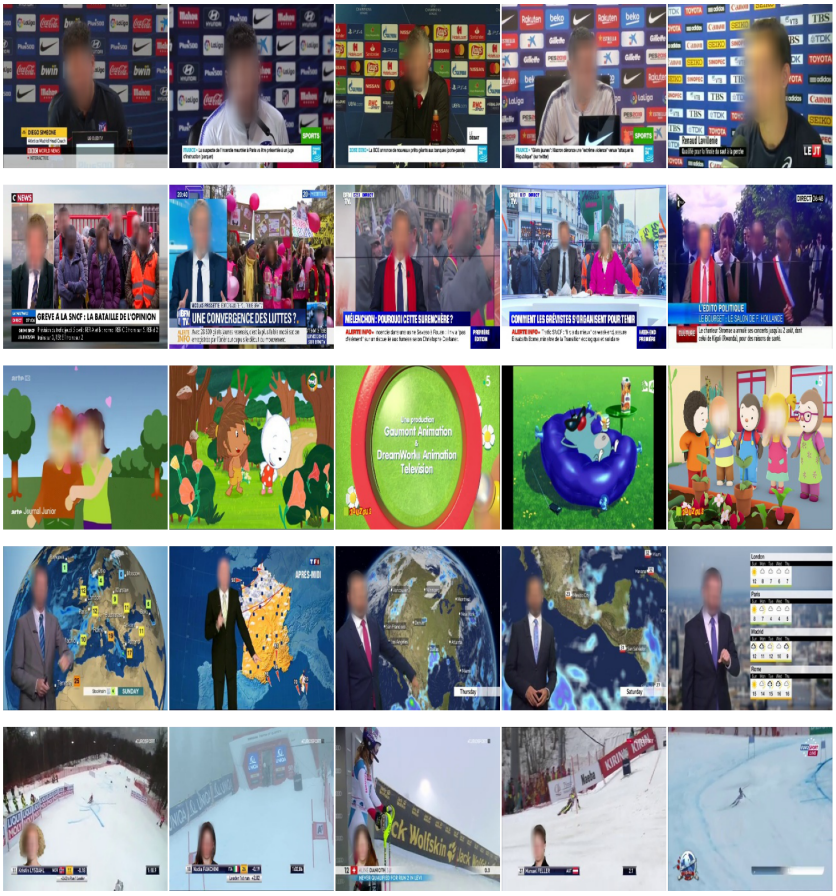

Fig. 6. Sample images and their nearest neighbors in the test set. The queries are on the left column, and the nearest neighbors are then displayed from the closest ones on the left to the furthest on the right.

It is, to our knowledge, the largest dataset available to learn a visual context descriptor that is specific to TV programs. We further present an original approach for visual context similarity metric learning specifically designed for identifying persons in TV programs. We show that our resulting neural network model can be used to effectively retrieve frames from semantically similar TV programs, and that it can be combined with state-of-the-art facial feature descriptors to improve the performance of a face verification task when such a visual context is available. We believe that the performance on the face verification tasks could be further improved with a suitable feature fusion strategy for facial descriptors and visual context descriptors.

References

1. Daudpota, S.M., Muhammad, A., Baber, J.: Video genre identification using clustering-based shot detection algorithm. Signal, Image and Video Processing **13**(7), 1413–1420 (2019)
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Conference on computer vision and pattern recognition. pp. 770–778 (2016)
3. Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In: Workshop on faces in'Real-Life'Images: detection, alignment, and recognition (2008)
4. Huang, Q., Xiong, Y., Lin, D.: Unifying identification and context learning for person recognition. In: Proc. of the Conference on Computer Vision and Pattern Recognition. pp. 2217–2225 (2018)
5. Kumar, N., Berg, A., Belhumeur, P.N., Nayar, S.: Describable visual attributes for face verification and image search. Transactions on Pattern Analysis and Machine Intelligence **33**(10), 1962–1977 (2011)
6. Petit, T., Letessier, P., Duffner, S., Garcia, C.: Unsupervised learning of co-occurrences for face images retrieval. In: International Conference on Multimedia in Asia. MMAsia '20, ACM (2021)
7. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proc. of the conference on computer vision and pattern recognition. pp. 815–823 (2015)
8. Smilkov, D., Thorat, N., Kim, B., Viégas, F., Wattenberg, M.: Smoothgrad: removing noise by adding noise. arXiv preprint arXiv:1706.03825 (2017)
9. Smirnov, E., Melnikov, A., Novoselov, S., Luckyanets, E., Lavrentyeva, G.: Doppelganger mining for face representation learning. In: Proc. of the International Conference on Computer Vision Workshops. pp. 1916–1923 (2017)
10. Smirnov, E., Melnikov, A., Oleinik, A., Ivanova, E., Kalinovskiy, I., Luckyanets, E.: Hard example mining with auxiliary embeddings. In: Proc. of the Conference on Computer Vision and Pattern Recognition Workshops. pp. 37–46 (2018)
11. Stone, Z., Zickler, T., Darrell, T.: Autotagging facebook: Social network context improves photo annotation. In: 2008 computer society conference on computer vision and pattern recognition workshops. pp. 1–8. IEEE (2008)
12. Varghese, J., Nair, K.R.: A novel video genre classification algorithm by keyframe relevance. In: Information and Communication Technology for Intelligent Systems, pp. 685–696. Springer (2019)
13. Wang, C., Zhang, X., Lan, X.: How to train triplet networks with 100k identities? In: International Conference on Computer Vision Workshops. pp. 1907–1915 (2017)
14. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: 2010 computer society conference on computer vision and pattern recognition. pp. 3485–3492. IEEE (2010)
15. Zhang, N., Paluri, M., Taigman, Y., Fergus, R., Bourdev, L.: Beyond frontal faces: Improving person recognition using multiple cues. In: Proc. of the Conference on Computer Vision and Pattern Recognition. pp. 4804–4813 (2015)
16. Zheng, L., Idrissi, K., Garcia, C., Duffner, S., Baskurt, A.: Triangular similarity metric learning for face verification. In: 11th International Conference and Workshops on Automatic Face and Gesture Recognition (FG). vol. 1, pp. 1–7. IEEE (2015)
17. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. Transactions on Pattern Analysis and Machine Intelligence (2017)