Video Analysis Based on Human Pose for Unsupervised Summarization and Retrieval

Conference Paper · September 2019

DOI: 10.1109/CBMI.2019.8877437

6 authors, including:

Carlos Santiago
Technical University of Lisbon
22 PUBLICATIONS **106** CITATIONS
SEE PROFILE

João Carvalho
Technical University of Lisbon
7 PUBLICATIONS **42** CITATIONS
SEE PROFILE

Joao Costeira
Technical University of Lisbon
92 PUBLICATIONS **2,191** CITATIONS
SEE PROFILE

VIDEO ANALYSIS BASED ON HUMAN POSE FOR UNSUPERVISED SUMMARIZATION AND RETRIEVAL

C Santiago[*], *DM Alves*[*], *BQ Ferreira*[*], *J Carvalho*[*], *A Messina*[†], *JP Costeira*[*]

[*]Institute for Systems and Robotics, LARSyS, Instituto Superior Técnico, Univ. Lisboa, Portugal
[†]RAI – Centre for Research and Technological Innovation, Turin, Italy

ABSTRACT

Finding good representations for videos is becoming increasingly more important to enable an efficient analysis and comparison, with potential applications in sports, surveillance, news, or web services. This paper proposes a new representation of videos based on human pose. Rather than looking at conventional features, our method relies only on human pose detections to characterize the video. This approach provides a powerful tool for the efficient analysis of videos of human activities, particularly for video summarization and retrieval. We evaluate the proposed representation on the following tasks: 1) computing video statistics, such as the main poses and viewpoint preferences; 2) partitioning videos into a collection of short clips that will compose the video summary; and 3) retrieving frames or scenes with specific poses from videos. Results show that the proposed approach is able to successfully perform these tasks.

Index Terms— Video analysis, video summarization, retrieval, pose detection, matrix completion

1. INTRODUCTION

The amount of videos captured, stored, and shared over the last decades has increased tremendously. Finding a good representation for this type of data is crucial to allow an efficient analysis of its content. This is shown to be particularly relevant for video summarization and retrieval, two tasks with a manifold of applications, such as sports, surveillance, news, or multimedia web search [1, 2].

Most state-of-the-art video analysis methods rely on deep learning architectures to both extract relevant features from the video frames and perform the desired task. For video summarization, typical approaches transform a full video into an short summary, either composed of the most relevant images (*keyframe summary*) or short clips (*video skimming*). When labeled training data is available, this may be viewed as the supervised learning problem of finding which video frames should be considered for the summary [3]. In this case, current state-of-the-art methods [4, 5] perform this task using recurrent neural networks, such as long short-term memory [6]. However, in many cases, labeled data is not available and thus video summarization has to be performed in an unsupervised setting. In these situations, this task is usually addressed as a keyframe selection problem using conventional image features [7, 8]. Specifically, [9] uses conventional "shallow" features (saliency) to predict people and object importance for the application of egocentric video summarization. Simple color features are also used in [10], where a specialized hierarchical clustering algorithm leveraging the temporal order and sequential nature of videos is proposed. As mentioned, more recently, features extracted from deep neural networks have become the *de facto* image descriptors. In [11], this type of features is combined with reinforcement learning to select the video frames, while [8] resorts to adversarial networks to identify which frames best represent the whole video.

For both image and video retrieval, current state-of-the-art approaches rely on approximate nearest neighbour algorithms, for a faster search [13], based on hashing strategies [14], which involve projecting a high-dimensional representation into compact, low-dimensional binary codes, to allow high computational efficiency and low memory requirements [15]. The image and video representations are based on deep features [16, 17], due to their discriminative power. However,

Fig. 1. Output of the OpenPose [12] algorithm for an image of a cycling race.

This work was supported by FCT through grants [PD/BD/114430/2016], [PD/BD/114429/2016] and [UID/EEA/50009/2013], and EU Horizon 2020 project MULTIDRONE (No 731667). The Titan Xp used for this research was donated by the NVIDIA Corporation.

978-1-7281-4673-7/19/$31.00

applying generic feature extractors may lead to poor results, as unbalanced feature relevances may cause hashing-based methods to yield lower performances [18, 19].

In this paper, we propose a new approach to the analysis of videos of human activities, such as sports events. Instead of relying on conventional image features (either shallow or deep), our approach uses only information about the poses of people in the scene. This way, our method takes as input more high-level, meaningful and low-dimensional information, which allows for a simpler approach and more interpretable reasoning than most state-of-the-art approaches. In particular, the videos are pre-processed using a pose estimation algorithm [12], which converts each frame (2D image) into a collection of keypoints from all the detected humans in the image, as shown in Fig. 1. Compared to other approaches, our strategy automatically filters out frames with no humans (e.g., landscapes). Furthermore, with this new representation, we are able to perform unsupervised video summarization and retrieval of specific frames or short clips of a scene based on a query image, for videos portraying any kind of human activity.

The proposed method is evaluated in the context of videos of cycling races on the tasks of: 1) identifying keyposes and viewpoints; 2) partitioning videos into short clips for unsupervised summarization; and 3) retrieval of specific frames or clips. We show that this framework is able to identify and select scenes of these events in a very efficient manner. Therefore, we deem this to be a powerful tool for directors to supervise broadcasts, for the post-editing staff to produce summaries and highlights, and for multimedia web search and retrieval.

2. PROPOSED APPROACH

This section describes the proposed methodology to represent videos of human activities. We assume that a video is a collection of scenes, each depicting humans viewed from a specific perspective, thus taking on a particular pose with respect to the camera. From a high level perspective, our approach first converts videos to a collection of detected human poses, which includes: 1) detecting all human poses in the video and 2) performing matrix completion to recover missing keypoints in the poses. Then, using this new representation, we can determine the keyposes by performing an unsupervised clustering of all detected poses. Finally, we can perform several common video analysis tasks, such as a statistical analysis, video summarization, and retrieval. Each of these steps are detailed below.

2.1. Pose Representation

As a first step, we start by detecting human poses on all video frames to convert the video into a collection of poses. Each pose is described by a set of keypoints in the human body (e.g., joints, head, etc). Clutter and occlusion are very common in this type of videos, which means that some of the keypoints may not be detected, jeopardizing further analysis of the video. We explicitly deal with this issue by applying a matrix completion algorithm that estimates the position of the missing keypoints. After converting the videos, we use this new representation to analyze them and extract relevant information.

2.1.1. Pose Estimation

Pose estimation is, in itself, a complex computer vision problem that has been extensively addressed in the literature in recent years [20, 12, 21, 22]. We propose to use the state-of-the-art pose detection OpenPose [12], which provides a 2D representation, based on keypoints, of all humans detected in each frame (see example in Fig. 1). We select this method since it is known to be efficient for images with multiple people and perform well in highly cluttered and occluded scenes. By applying the OpenPose algorithm, the video is converted into a collection of poses.

Formally, let $\mathbf{x}^k = \left[\mathbf{x}_1^k, \mathbf{x}_2^k, \ldots, \mathbf{x}_N^k\right] \in \mathbb{R}^{2N}$ be the vector representing a specific pose k, described by N keypoints (in OpenPose, $N = 18$), where $\mathbf{x}_i^k \in \mathbb{R}^2$ is the position of the i-th keypoint in the image. We denote

$$\mathbf{X}_0 = [\mathbf{x}^1, \ldots, \mathbf{x}^M] \tag{1}$$

as the matrix collecting, column-wise, all the M poses detected in the video. As previously mentioned, some entries in $\mathbf{X}_0$ are often unknown, due to clutter and occlusions. By definition, missing entries in $\mathbf{X}_0$ are replaced by zeros. In practice, to analyze a video, we need the complete matrix $\mathbf{X}$, in which all the entries (i.e., the position of all the keypoints) are known. To obtain $\mathbf{X}$ from $\mathbf{X}_0$, we use the matrix completion approach described in the following section.

2.1.2. Recovering Missing Keypoints

The poses with missing keypoint detections make further analysis of the video harder. To deal with this issue, we can apply several state-of-the-art methods that are able to estimate the missing entries [23, 24, 25, 26]. For this application, we found Subspace Segmentation by Successive Approximations (SSSA) [26] to provide more reliable completion than the other tested approaches [23, 24, 25]. This method performs completion under the assumption that the data lies in the union of linear subspaces and that any point (pose) can be represented as a linear combination of other points (poses) on the data.

Let Ω be the set of indices of the known entries in $\mathbf{X}_0$. SSSA finds the complete matrix, $\mathbf{X} \in \mathbb{R}^{2N \times M}$, by solving

the following optimization problem

$$\{\mathbf{X}, \mathbf{C}, \mathbf{E}\} = \min_{\mathbf{X},\mathbf{C},\mathbf{E}} \quad \|\mathbf{C}\|_1 + \frac{\lambda}{2}\|\mathbf{E}\|_F^2 \qquad (2)$$
$$\text{s.t.} \quad \mathbf{X} = \mathbf{X}\mathbf{C} + \mathbf{E}$$
$$P_\Omega(\mathbf{X}) = P_\Omega(\mathbf{X}_0)$$
$$diag(\mathbf{C}) = 0,$$

where $\mathbf{C}$ is the matrix of coefficients, $\mathbf{E}$ is the matrix of reconstruction errors, and $P_\Omega(\cdot)$ indicates the Ω entries of a matrix. Note that minimizing the ℓ_1-norm of $\mathbf{C}$ results on a sparse solution in which (ideally) a given pose is represented as a linear combination of only a few similar poses.

Problem (2) is the generalization of the Sparse Subspace Clustering [27] for the case with incomplete data. However, due to the product between $\mathbf{X}$ and $\mathbf{C}$, the problem becomes non-convex. To solve it we follow [26] which proposes an alternate algorithm that takes advantage of the Alternating Direction Method of Multipliers [28].

The subsequent video analysis is performed using this new representation of the video, $\mathbf{X}$, and the corresponding time stamps (the frame in which each pose appears), instead of the original data.

2.2. Pose Clustering

We could take advantage of the above formulation to apply Spectral Clustering [29] on an affinity matrix built from matrix $\mathbf{C}$. However, this method is not scalable with respect to the number of detected poses, making this strategy infeasible for this context. Therefore, we follow a different clustering strategy to find the keyposes in the videos.

The clustering methodology we adopt in this work is the k-means++ algorithm [30], based on the Euclidean distance ($||\mathbf{x}^i - \mathbf{x}^j||_2$ for poses i and j). Since this distance is not invariant to translation nor scaling, each pose is translated to have the head keypoint centered in the origin, and scaled based on the average length of the upper arms. k-means++ finds a specific number of pose clusters, K, given the detected poses in $\mathbf{X}$. The number of clusters is unknown a priori, but an expected number can be roughly estimated for each type of video.

3. RESULTS

In this section, we present and discuss the results of applying our proposed approach. We evaluate our approach on the application of cycling races video analysis, which is one of the main sports event covered in the scope of the MULTIDRONE project, using video broadcasts from RAI's Giro D'Italia 2017[1]. Specifically, we use 27 minutes of video from a one stage of the race as training, to learn the main poses and perform video partitioning. Then, we evaluate our method in terms of statistical analysis, video partitioning and search and retrieval, on two test videos of different stages with a total duration of 17 minutes.

[1] from https://multidrone.eu/multidrone-public-dataset/

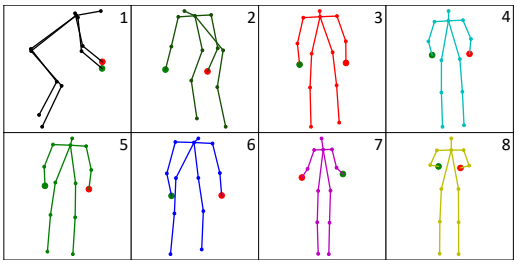

Fig. 2. Centroids of pose clusters obtained for the training video. Clusters 1 to 7 correspond to cyclists viewed from both sides (1-2), front (3-6) and back (7), while cluster 8 corresponds to spectators.

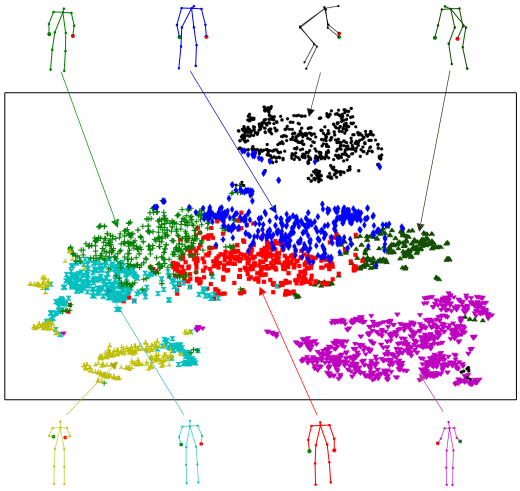

Fig. 3. t-SNE visualization of the clustering obtained for test video 1

3.1. Pre-processing

When converting videos to the collection of detected poses, we only take as input OpenPose detections that have at least one keypoint detected with a confidence score (output by OpenPose) greater than 0.9, filtering out possible false positives or inaccurate poses. Additionally, we discard the 4 facial keypoints (eyes and ears) as we considered them not to be relevant for discriminating the human poses. Hence, each $\mathbf{x}^k$ has $N = 14$ keypoints, and the matrix collecting all poses, $\mathbf{X}$, has dimensions $28 \times M$.

3.2. Clustering

We empirically found $K = 8$ to be the best value in the range of $\{4, \ldots, 10\}$, accounting for the most relevant cy-

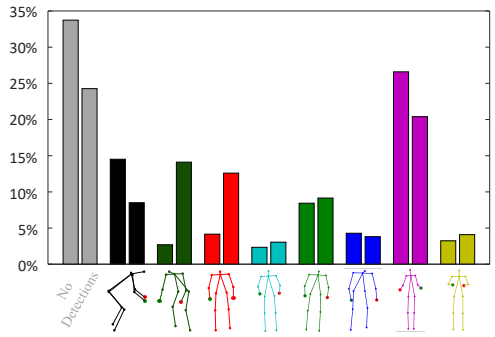

Fig. 4. Histogram of the frames of the two test videos assigned (by majority voting) to each cluster found for the training video (see Fig. 2). The first columns correspond to frames without pose detections.

clists poses, as well as outliers such as spectators and other passersby. Applying our pose clustering to the training video, which has a total of 39K detected poses (from which aprox. 37% had missing keypoints), produces the cluster centroids depicted in Fig. 2 (left hand is red and right hand is green). From these results we can see that our approach can distinguish scenes with cyclists viewed from both sides, from the back and front, and within the latter cyclists leaning with different angles. The last centroid (8) corresponds to spectators as it includes poses of people clapping and cheering.

Since the clustering method is performed with data points of dimension 28, we apply the dimensionality reduction method t-Stochastic Neighbor Embedding (t-SNE) [31] to visualize the resulting clusters in 2D. Fig. 3 depicts the t-SNE view of the detected poses in test video 1, where the color matches the corresponding cluster. This visualization shows that spectators are mostly separated (bottom) from the cyclists. Also, side and back views of cyclists are clearly separated from the front views.

3.3. Video Analysis

Once the videos have been converted to the pose representation, and the keyposes have been identified, further analysis of what are the most important viewpoints and scene types can be performed. By assigning a label to each frame (based on the predominant pose per frame), it is also possible to: 1) compute video statistics of the most frequent scene types; and 2) partition the video into short clips of the same type. Broadcasts of cycling races frequently alternate between showing cyclists and other types of scenes, such as aerial views, scenery, spectators and close-ups on specific actions. Since OpenPose only detects poses within a limited range of scales, we know beforehand that it will not provide detections for all scenes. We also use this information to automatically distinguish between these types of scenes.

3.3.1. Video Statistics

The histograms in Fig. 4 show the frequency of each type of scene for each of the two test videos. It is interesting to note that, as expected, spectators related views are the least frequent ones for both stages and that the different cyclists views are almost equally frequent (except for the side views, as one of the video stages depicts nearly no cyclists moving from right to left). We can also verify that the majority of the frames do not show any poses. These correspond to landscapes scenes, aerial views of the race and close ups on the cyclists, all of which yield no pose detections. This analysis can be used for video summarization, since more frequent keyposes are good proxies for the selection of frames for the *keyframe summary* of the video. These histograms also provide a good representation of the directors preferences for that specific event.

3.3.2. Video Partitioning and Summarization

Assigning a label to each frame also allows videos to be partitioned into different scenes. Applying a median filter enforces smoothness in the temporal labeling. Partitioning results are shown for test video 1 in Fig. 5. This figure shows that the proposed method is able to identify segments of the video in which the cyclists are viewed from a specific perspective. It is also clear that the video transitions between specific viewpoints and scenes in which the cyclists are either too far (as shown by the fourth example) or too close, making the OpenPose algorithm unable to detect poses.

The video partitions obtained with this strategy can be used to search and detect scenes based on the requested pose. Furthermore, it is possible to perform video summarization by either: 1) generating a collection of short clips from each partition (*video skimming*); or 2) by selecting one frame from each partition (*keyframe summary*), as shown by the 6 examples in Fig. 5.

3.3.3. Search and Retrieval of Specific Poses

The pose representation also allows searching for specific poses in videos. As an example, we use test video 2 to search for poses similar to the keyposes 1, 2 and 3 shown in Fig. 2. The retrieved results are shown in Fig. 6, which depicts the 3 different frames found to have the most similar poses to each of the keyposes.

These results show that this approach is able to retrieve frames with specific pose detections from different videos. This is also very useful in other applications, such as in surveillance, to identify humans in abnormal or suspicious activities. Additionally, by assigning a label to each frame, as represented in Fig. 5, we can also perform a image to video retrieval, where an input image query is used to retrieve short clips within videos depicting scenes with similar poses.

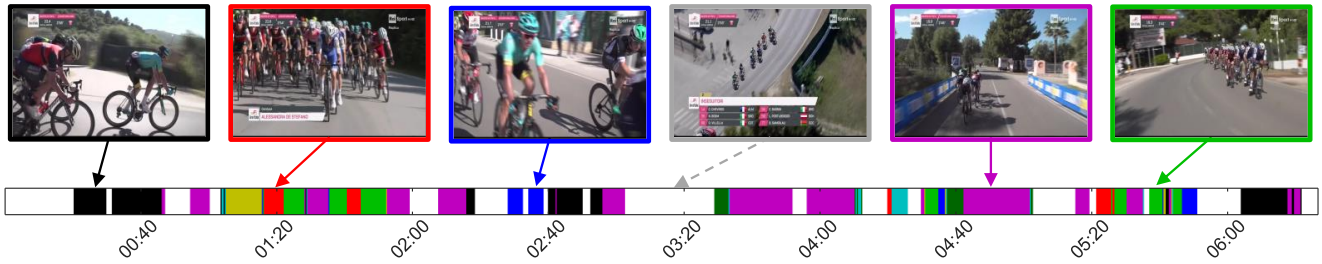

Fig. 5. Video partitioning based on the detected poses. The colors correspond to the clusters shown in Fig. 2. Regions with no color (white) correspond to frames with no detections.

Fig. 6. Examples of 3 frames from test video 2 similar to the keyposes 1, 2 and 3 shown in Fig. 2, respectively.

4. CONCLUSIONS

We propose a new representation for the automatic analysis of videos of human activities. This representation is based on high-level features based on detected human poses, which allows an efficient analysis of the video contents. The proposed representation was evaluated on the following tasks: 1) computing video statistics, such as the main poses and viewpoint preferences; 2) partitioning videos into a collection of short clips, each depicting a specific scene or viewpoint, from which we perform unsupervised video summarization; and 3) retrieving specific frames or clip within videos based on a query image. Results show that the proposed approach is able to handle these tasks very well, making this framework a step forward in the automatic analysis of this type of videos.

Future work includes using more robust metrics to assign keypose-labels to video frames, such as considering the distribution of detect poses (instead of the most frequent pose), as well as using temporal information, as this will ensure that the video partitions and corresponding summaries are more reliable and consistent. Similarly, the ranking metric used to perform the retrieval task should account not only for the similarity between individual poses, but also the pose distribution and temporal context. This will enforce the retrieved frames to correspond to parts of the video with a similar scene, instead of simply containing a pose similar to the input query.

5. REFERENCES

[1] J. Almeida, N. J. Leite, and R. d. S. Torres, "Vison: Video summarization for online applications," *Pattern Recognition Letters*, vol. 33, no. 4, pp. 397–409, 2012.

[2] K. Zhang, K. Grauman, and F. Sha, "Retrospective encoders for video summarization," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2018, pp. 383–399.

[3] B. Gong, W. L. Chao, K. Grauman, and F. Sha, "Diverse sequential subset selection for supervised video summarization," in *Advances in Neural Information Processing Systems*, 2014, pp. 2069–2077.

[4] M. Rochan, L. Ye, and Y. Wang, "Video Summarization Using Fully Convolutional Sequence Networks," *arXiv preprint arXiv:1805.10538*, 2018.

[5] K. Zhang, W. L. Chao, F. Sha, and K. Grauman, "Video summarization with long short-term memory," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 766–782.

[6] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[7] Z. Tian, J. Xue, X. Lan, C. Li, and N. Zheng, "Object segmentation and key-pose based summarization for motion video," *Multimedia Tools and Applications*, vol. 72, no. 2, pp. 1773–1802, 2014.

[8] B. Mahasseni, M. Lam, and S. Todorovic, "Unsupervised video summarization with adversarial LSTM networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, 2017.

[9] Y. J. Lee, J. Ghosh, and K. Grauman, "Discovering important people and objects for egocentric video summarization," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1346–1353.

[10] K. M. Mahmoud, N. M. Ghanem, and M. A. Ismail, "Unsupervised video summarization via dynamic modeling-based hierarchical clustering," in *International Conference on Machine Learning and Applications*, vol. 2, 2013, pp. 303–308.

[11] K. Zhou, Y. Qiao, , and T. Xiang, "Deep reinforcement learning for unsupervised video summarization with," in *AAAI Conference on Artificial Intelligence*, 2018.

[12] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7291–7299.

[13] K. Li and J. Malik, "Fast k-nearest neighbour search via dynamic continuous indexing," in *International Conference on Machine Learning*, 2016, pp. 671–679.

[14] H. Liu, R. Wang, S. Shan, and X. Chen, "Deep supervised hashing for fast image retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2064–2072.

[15] Z. Lin, G. Ding, J. Han, and J. Wang, "Cross-view retrieval via probability-based semantics-preserving hashing," *IEEE Transactions on Cybernetics*, vol. 47, no. 12, pp. 4342–4355, 2017.

[16] A. Babenko and V. Lempitsky, "Aggregating deep convolutional features for image retrieval," *arXiv preprint arXiv:1510.07493*, 2015.

[17] A. S. Razavian, J. Sullivan, S. Carlsson, and A. Maki, "Visual instance retrieval with deep convolutional networks," *ITE Transactions on Media Technology and Applications*, vol. 4, no. 3, pp. 251–258, 2016.

[18] Z. Dong, S. Jia, T. Wu, and M. Pei, "Face video retrieval via deep learning of binary hash representations," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[19] G. Wu, J. Han, Y. Guo, L. Liu, G. Ding, Q. Ni, and L. Shao, "Unsupervised deep video hashing via balanced code for large-scale video retrieval," *IEEE Transactions on Image Processing*, vol. 28, no. 4, pp. 1993–2007, April 2019.

[20] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, "Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 561–578.

[21] R. A. Güler, N. Neverova, and I. Kokkinos, "Densepose: Dense human pose estimation in the wild," *arXiv preprint arXiv:1802.00434*, 2018.

[22] X. B. Peng, A. Kanazawa, J. Malik, P. Abbeel, and S. Levine, "Sfv: Reinforcement learning of physical skills from videos," *arXiv preprint arXiv:1810.03599*, 2018.

[23] J. F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[24] Z. Wen, W. Yin, and Y. Zhang, "Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm," *Mathematical Programming Computation*, vol. 4, no. 4, pp. 333–361, 2012.

[25] J. Fan and T. W. Chow, "Sparse subspace clustering for data with missing entries and high-rank matrix completion," *Neural Networks*, vol. 93, pp. 36–44, 2017.

[26] J. Carvalho, M. Marques, and J. P. Costeira, "Recovery of Subspace Structure from High-Rank Data with Missing Entries," in *IEEE International Conference on Image Processing (to appear)*, 2019.

[27] E. Elhamifar and R. Vidal, "Sparse subspace clustering," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 2790–2797.

[28] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein *et al.*, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine learning*, vol. 3, no. 1, pp. 1–122, 2011.

[29] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in neural information processing systems*, 2002, pp. 849–856.

[30] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, 2007, pp. 1027–1035.

[31] L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.