A Fixed-point Estimation Algorithm for Learning The Multivariate GGMM: Application to Human Action Recognition

Fatma Najar*, Sami Bourouis†‡, Nizar Bouguila§ and Safya Belghith *
*Université de Tunis El Manar, ENIT, Lab: Robotique Informatique et Systèmes Complexes, Tunis, Tunisie
Email: fatma.najjar@enit.utm.tn, safya.belghith@enit.utm.tn
†Université de Tunis El Manar, ENIT, Lab: Signal, Image et Technologies de l'Information, Tunis, Tunisie
‡Taif University, Taif, Saudi Arabia
Email:s.bourouis@tu.edu.sa
§The Concordia Institute for Information Systems Engineering, Concordia University, Montreal, Canada.
Email: nizar.bouguila@concordia.ca

***Abstract*—Multivariate generalized Gaussian distribution has been an attractive solution to many signal and image processing applications. Therefore, efficient estimation of its parameters is of significant interest for a number of research problems. The main contribution of this paper is to develop a fixed-point estimation algorithm for learning the multivariate generalized Gaussian mixture model's parameters (MGGMM). A challenging application that concerns Human action recognition is deployed to validate our statistical framework and to show its merits.**

***Index Terms*—Multivariate Generalized Gaussian distribution, mixture models, Fixed-point estimation, Expectation-maximization, Human action recognition.**

I. Introduction

Human activity recognition has been an active research topic in the last few years for several applications related to computer vision and pattern recognition [1], [2], [3], [4]. Actions recognition plays a significant role in video surveillance systems, human-to-human interaction, robotics, health care activities, etc. However, human actions recognition is still a difficult problem due to the high complexity of actions as well as their variability, and the difficulties caused by the presence of background clutter and illumination changes. In literature, different statistical learning models have been used for Human action recognition [1], [2]. Recently, a new learning framework for human action representation has been proposed in [5]. It uses a convolutional neural networks (CNNs) and a linear dynamical system (LDS) to represent both spatial and temporal structures of actions in video sequences. In [6] an unsupervised learning method for human action categorization was developed based on a probabilistic latent semantic analysis (pLSA) model and latent Dirichlet allocation (LDA). Gaussian mixture models (GMM) are used with success for recognizing actions. Indeed, a classification method that uses GMM, Hidden Markov Model (HMM) and skeleton features was proposed in [7]. Another adaptive GMM has been developed in [8] to classify human actions based on background segmentation and visual information. In [9], authors apply both GGM and regression technique to build a computational models of human motion learned from human examples which is used also for human action classification purposes. On the other hand, there has been a substantial growth for developing mixture models using generalized Gaussian mixture models (GGMMs) since they provide a flexibility to fit the shape of the data better than the Gaussian distribution. GGMMs have been widely used in image and video classification and segmentation [10], [11], [12], texture image analysis [13], text independent speaker identification [14] and image denoising [15]. For the case of multivariate analysis, the previous cited works and many others used only the diagonal covariance matrix and ignore the richness of the full covariance matrix for calculation simplicity purposes. Their justification is based also on the fact that the features of observed data are independent. However, features are not always independent and for several applications, like the case of human activities, data can be correlated. Therefore, to deal with such problem, some covariance matrix estimators have been developed such as the so-called Fixed-point based estimator [16]. It is used to estimate the full covariance matrix and the shape parameter of a zero-multivariate generalized Gaussian Distribution. Thus, we are mainly motivated in this work by investigating this line of research by introducing, for the first time, a Fixed-point estimation algorithm for learning the multivariate GGMM. The whole algorithm is driven by the famous Expectation-Maximization algorithm and it is applied for Human activity recognition application.

The remainder of this paper is organized as follows: in section 2, we review estimation method for the case of multivariate generalized Gaussian distribution. Then, we present in section 3 our proposed multivariate generalized Gaussian mixture model. Section 4 is dedicated for experiments and finally, we end this paper with some conclusions and future directions.

II. Multivariate Generalized Gaussian Distribution

Multivariate generalized Gaussian distributions (MGGDs) belong to the family of elliptical distributions originally introduced in [17]. MGGDs are characterized by a mean vector, a covariance matrix and a shape parameter. They are defined by their probability density functions [18] as follow:

$$p(X|\Sigma;\beta;\mu) = \frac{\Gamma(\frac{K}{2})}{\pi^{\frac{K}{2}}\Gamma(\frac{K}{2\beta})2^{\frac{K}{2\beta}}}\frac{\beta}{m^{\frac{K}{2}}|\Sigma|^{\frac{1}{2}}} \times \tag{1}$$

$$exp\Big[-\frac{1}{2m^{\beta}}((X-\mu)^{T}\Sigma^{-1}(X-\mu))^{\beta}\Big] \tag{2}$$

where $X \in R^{K}$, m is the scale parameter, $\beta > 0$ is the shape parameter and Σ is a $K \times K$ symmetric positive definite matrix, called the covariance matrix. If $\beta = 1$, the MGGD is equivalent to the multivariate Gaussian distribution. The shape parameter β controls the peakedness and the spread of the distribution. If $\beta < 1$, the distribution is more peaky than Gaussian with heavier tails, and if $\beta > 1$, it is less peaky with lighter tails.

A. Fixed-Point estimation method

One of the well defined estimation techniques of the MGGD parameters is the so-called "Fixed Point method" [16]. More precisely, for any shape parameter belonging to [0,1], the Maximum Likelihood estimator of the covariance matrix exists and is unique. The existence was proved by showing that the profile likelihood is positive, bounded in the set of symmetric positive definite matrices and equals zero on the boundary of this set. Regarding the uniqueness, it was proved that for any initial symmetric positive definite matrix, the sequence of matrices satisfying a fixed point equation converging to the unique maximum of this profile likelihood. Afterwards, an iterative algorithm based on a Newton-Raphson technique is then applied to compute the MLE of the shape parameter. Let $(x_1, x_2, ..., x_T)$ be a random sample of T observation vectors of dimension K, drawn from a zero mean MGGD with scatter matrix $M = m\Sigma$; m is the scale parameter, and β is the shape parameter. The ML estimators of m, β and Σ are defined by :

$$\hat{\Sigma}_{k+1} = f(\Sigma_k) \tag{3}$$

where

$$f(\Sigma) = \sum_{i=1}^{T} \frac{K}{u_i + u_i^{1-\beta}\sum_{i\neq j} u_j^{\beta}} x_i x_i^{T}, \tag{4}$$

$$\hat{m} = \Big[\frac{1}{T}\sum_{i=1}^{T}(u_i)^{\beta}\Big]^{\frac{1}{\beta}}, \tag{5}$$

Where $u_i = x_i^{T}\Sigma^{-1}x_i$

$$\hat{\beta}_{k+1} = \hat{\beta}_k - \frac{\alpha(\hat{\beta}_k)}{\alpha'(\hat{\beta}_k)} \tag{6}$$

where

$$\alpha(\beta) = \frac{KT}{2\sum_{i=1}^{T} u_i^{\beta}} \sum_{i=1}^{T} \Big[u_i^{\beta} ln(u_i)\Big] - \frac{KT}{2\beta}\Big[\psi\Big(\frac{K}{2\beta}\Big) + ln(2)\Big] \\ -T - \frac{KT}{2\beta} ln\Big(\frac{\beta}{KT}\sum_{i=1}^{T} u_i^{\beta}\Big) \tag{7}$$

Where ψ is the digamma function.

III. Multivariate Generalized Gaussian Mixture Models

In the previous section, we have presented the estimation of the MGGD's parameters for the case of a single distribution. Now, we develop in this section a Fixed-point estimation algorithm for learning the multivariate generalized mixture model. The general form of a mixture model is given as:

$$f(X|\Theta) = \sum_{j=1}^{M} p_j p(X|\Sigma_j;\beta_j;\mu_j) \tag{8}$$

where $\forall j, p_j \geq 0; \sum_j p_j = 1$, each Θ_j is the set of parameters of the j^{th} component and $\Theta = \{\Sigma_1, ..., \Sigma_M; \beta_1, ..., \beta_M; \mu_1, ..., \mu_M\}$ denotes the full parameter set. For parameter's estimation, one of the most popular used approach is the Maximum Likelihood method which is used in conjunction with the Expectation-Maximization (EM) algorithm [19]. The log-likelihood is given by:

$$log(X|\Theta) = \sum_{i=1}^{T} log(f(X_i|\Theta)) \tag{9}$$

The proposed algorithm is summarized as follow:

Algorithm 1 FP-MGGMM

```
1) Initialization step : Initializing model's parameters with the k-means algorithm
   followed by the method of moment applied to each cluster.
2) Repeat until convergence of the log-likelihood :
   • Expectation step : Computing responsibilities
         $p(j|X_i) = \frac{p_j p(X_i|\Sigma_j;\beta_j;\mu_j)}{\sum_{m=1}^{M} p_m p(X_i|\Sigma_m;\beta_m;\mu_m)}$   (10)
   • Maximization step
     – Mean estimation
         $\hat{\mu}_j = \frac{\sum_{i=1}^{T} p(j|X_i)|X_i - \mu_j|^{\beta_j - 1} X_i}{\sum_{i=1}^{T} p(j|X_i)|X_i - \mu_j|^{\beta_j - 1}}$   (11)
     – Covariance estimation of each cluster : Normalizing the dataset ($X_n = X - \mu_j$),
       then evaluating the covariance matrix using equations 3 and 4.
     – Shape estimation : The shape parameter is determined using equation 6 and 7.
3) Assign each data point to the nearest cluster through the Bayes' rule.
```

IV. Experimental results

In this section, we apply our algorithm on a real challenging application namely the human action recognition. We evaluate the performance of our proposed mixture model denoted by FP-MGGMM and we compare the obtained results w.r.t those provided by the classic Gaussian mixture model and generalized Gaussian mixture model driven only by the diagonal covariance matrix. In our experiments, we adopt the following strategy (that involves three stages) to classify the human actions from images: feature extraction, image representation, and action classification. In feature extraction, we have used dense SIFT descriptors of 16×16 pixel patches computed over a grid with spacing of 8 pixels. Next, we apply the bag of words (BOW) technique [20] for classification purposes. The BOW allows to quantize the image features into visual words on the basis of the K-means algorithm. Thus, each image is represented as a frequency histogram over the V visual words. The last step is the application of a probabilistic Latent Semantic Analysis (pLSA) to the obtained histograms in order to represent each image by a D-dimensional vector where D is the number of latent aspects [21]. Finally, we classify the overall images to their right activities using our FP-MGGMM algorithm. We have considered publicly available datasets such as UIUC Sport Event dataset [22] and Stanford 40 Action dataset [23]. Those datasets are very challenging because most images have highly cluttered and diverse background, and object classes are highly distinct. Some samples of images from these datasets are shown in Fig. 1 and Fig. 2 respectively. For the learning process, we randomly select 100 images in each class for training, and the remaining images for testing.

Fig. 1: Sample images from the UIUC sports event dataset.

Fig. 2: Sample images from the Stanford 40 Action dataset.

A comparative study between different methods (GMM, GGMM, and FP-MGGMM) is depicted in Table I. It represents the average classification accuracy rate for both UIUC and Stanford datasets. According to these results, we can see that the FP-MGGMM offers the highest average accuracy rate (it is about 34% for UIUC and 42% for Stanford) and it outperforms other models which assume that the dimensions of the observed data are independent. This means that the consideration of the full covariance matrix through the Fixed-point algorithm helps in improving the expected performances.

TABLE I: The average classification accuracy rate for different mixture models

Algorithm	UIUC dataset	Stanford dataset
GMM	30.52	34.80
GGMM	31.69	35.20
FP-MGGMM	**34.41**	**42.13**

In other words, better results are obtained when taking into account the full covariance matrix and not when using only the diagonal matrices (i.e considering only the standard deviation values). This interpretation leads to the following assumption: more features used in the covariance matrix to describe the actions, better classification performances can be obtained. We have evaluated also the impact of different visual vocabulary sizes on the classification accuracy for the proposed FP-MGGMM algorithm, as illustrated in Fig.3 and Fig.4. As we can see, the maximum classification rate is obtained with visual vocabulary sizes of 50 and 600 for UIUC and Stanford datasets respectively.

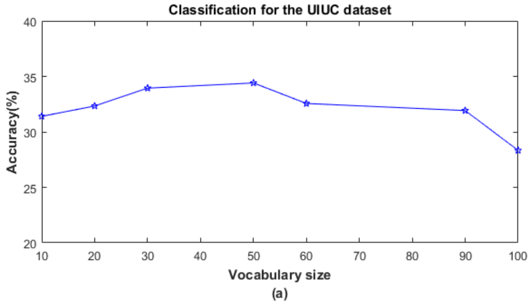

Fig. 3: Classification accuracy w.r.t vocabulary size for UIUC dataset.

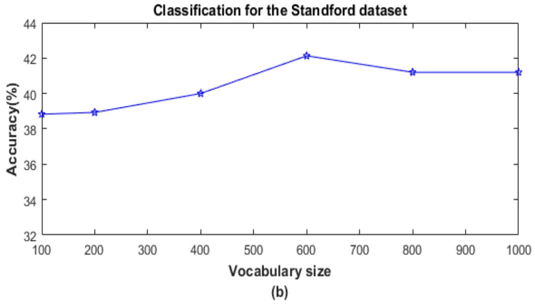

Fig. 4: Classification accuracy w.r.t vocabulary size for Stanford dataset.

We found also that the classification accuracy is affected by the choice of the number of aspects. As shown in Fig.5 and Fig.6, the optimal accuracy was obtained when the number of aspects was set to 8 for UIUC and 10 for Stanford dataset.

V. Conclusion

In this paper, we have presented a novel unsupervised Fixed-point estimation algorithm for learning the multivariate generalized Gaussian mixture model that uses the full

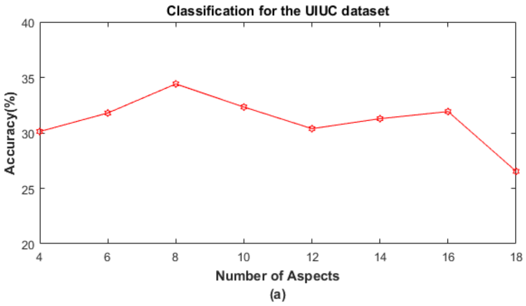

Fig. 5: Classification accuracy w.r.t the number of aspects for UIUC dataset.

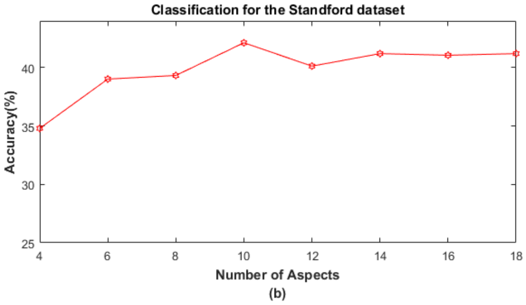

Fig. 6: Classification accuracy w.r.t the number of aspects for Stanford dataset.

covariance matrix. The proposed method is driven by the famous Expectation-Maximization algorithm and it is applied for Human activity recognition application. We evaluated the performance of the proposed framework through two publicly available datasets: UIUC Sport Event dataset and Stanford 40 Action. Obtained results are encouraging and show that our model outperforms the GMM and GGMM which are based only on the diagonal covariance matrix. Future work will focus on the improvement of obtained results by taking into account more relevant visual features like in [24], [25] and also by adopting a semi-supervised or a weak-supervised setting. We propose also to evaluate our model on video sequences classification.

References

[1] T. Subetha and S. Chitrakala, "A survey on human activity recognition from videos," in *Information Communication and Embedded Systems (ICICES), 2016 International Conference on*. IEEE, 2016, pp. 1–7.

[2] M. Vrigkas, C. Nikou, and I. A. Kakadiaris, "A review of human activity recognition methods," *Frontiers in Robotics and AI*, vol. 2, p. 28, 2015.

[3] W. Fan, H. Sallay, N. Bouguila, and J. Du, "Human action recognition using accelerated variational learning of infinite dirichlet mixture models," in *14th IEEE International Conference on Machine Learning and Applications, ICMLA 2015, Miami, FL, USA, December 9-11, 2015*, 2015, pp. 451–456.

[4] W. Fan and N. Bouguila, "A variational statistical framework for clustering human action videos," in *13th International Workshop on Image Analysis for Multimedia Interactive Services, WIAMIS 2012, Dublin, Ireland, May 23-25, 2012*, 2012, pp. 1–4.

[5] L. Zhang, Y. Feng, X. Xiang, and X. Zhen, "Realistic human action recognition: When cnns meet lds," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 1622–1626.

[6] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *International journal of computer vision*, vol. 79, no. 3, pp. 299–318, 2008.

[7] L. Piyathilaka and S. Kodagoda, "Gaussian mixture based hmm for human daily activity recognition using 3d skeleton features," in *Industrial Electronics and Applications (ICIEA), 2013 8th IEEE Conference on*. IEEE, 2013, pp. 567–572.

[8] Y. Dedeoglu, "Human action recognition using gaussian mixture model based background segmentation," in *Machine Learning Workshop, Bilkent University*, 2005.

[9] B. Bruno, F. Mastrogiovanni, A. Sgorbissa, T. Vernazza, and R. Zaccaria, "Human motion modelling and recognition: A computational approach," in *Automation Science and Engineering (CASE), 2012 IEEE International Conference on*. IEEE, 2012, pp. 156–161.

[10] M. S. Allili, N. Bouguila, and D. Ziou, "Online video foreground segmentation using general gaussian mixture modeling," in *2007 IEEE International Conference on Signal Processing and Communications*, 2007, pp. 959–962.

[11] ——, "Finite generalized gaussian mixture modeling and applications to image and video foreground segmentation," in *Computer and Robot Vision, 2007. CRV'07. Fourth Canadian Conference on*. IEEE, 2007, pp. 183–190.

[12] F. Najar, S. Bourouis, N. Bouguila, and S. Belguith, "A comparison between different gaussian-based mixture models," in *14th IEEE International Conference on. Computer Systems and Applications, Tunisia*. IEEE, 2017.

[13] K. N. Kumar, K. S. Rao, Y. Srinivas, and C. Satyanarayana, "Studies on texture segmentation using d-dimensional generalized gaussian distribution integrated with hierarchical clustering," *International Journal of Image, Graphics and Signal Processing*, vol. 8, no. 3, p. 45, 2016.

[14] V. Sailaja, K. Srinivasa Rao, and K. Reddy, "Text independent speaker identification with finite multivariate generalized gaussian mixture model and hierarchical clustering algorithm," *Int. Journal of Computer Applications*, vol. 11, no. 11, pp. 0975–8887, 2010.

[15] I. Channoufi, S. Bourouis, N. Bouguila, and K. Hamrouni, "Image and video denoising by combining unsupervised bounded generalized gaussian mixture modeling and spatial information," *Multimedia Tools and Applications*, Feb 2018. [Online]. Available: https://doi.org/10.1007/s11042-018-5808-9

[16] F. Pascal, L. Bombrun, J.-Y. Tourneret, and Y. Berthoumieu, "Parameter estimation for multivariate generalized gaussian distributions," *IEEE Transactions on Signal Processing*, vol. 61, no. 23, pp. 5960–5971, 2013.

[17] D. Kelker, "Distribution theory of spherical distributions and a location-scale parameter generalization," *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 419–430, 1970.

[18] S. Kotz, "Multivariate distributions at a cross-road," *Statistical distributions in scientific work*, vol. 1, pp. 247–270, 1975.

[19] C. Bishop, "Pattern recognition and machine learning (information science and statistics), 1st edn. 2006. corr. 2nd printing edn," *Springer, New York*, 2007.

[20] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Workshop on statistical learning in computer vision, ECCV*, vol. 1, no. 1-22. Prague, 2004, pp. 1–2.

[21] A. Bosch, A. Zisserman, and X. Muñoz, "Scene classification via plsa," *Computer Vision–ECCV 2006*, pp. 517–530, 2006.

[22] L.-J. Li and L. Fei-Fei, "What, where and who? classifying events by scene and object recognition," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. IEEE, 2007, pp. 1–8.

[23] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, and L. Fei-Fei, "Human action recognition by learning bases of action attributes and parts," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1331–1338.

[24] I. Channoufi, S. Bourouis, N. Bouguila, and K. Hamrouni, "Color image segmentation with bounded generalized gaussian mixture model and feature selection," *Accepted, to be appear in the 4th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP'2018)*, 2018.

[25] W. Fan, H. Sallay, N. Bouguila, and S. Bourouis, "A hierarchical dirichlet process mixture of generalized dirichlet distributions for feature selection," *Computers & Electrical Engineering*, vol. 43, pp. 48–65, 2015.