Melanoma Detection Algorithm Based on Feature Fusion

Catarina Barata [1], M. Emre Celebi [2] and Jorge S. Marques [1]

Abstract— **A Computer Aided-Diagnosis (CAD) System for melanoma diagnosis usually makes use of different types of features to characterize the lesions. The features are often combined into a single vector that can belong to a high dimensional space (early fusion). However, it is not clear if this is the optimal strategy and works on other fields have shown that early fusion has some limitations. In this work, we address this issue and investigate which is the best approach to combine different features comparing early and late fusion. Experiments carried on the datasets PH^2 (single source) and EDRA (multi source) show that late fusion performs better, leading to classification scores of Sensitivity = 98% and Specificity = 90% (PH^2) and Sensitivity = 83% and Specificity = 76% (EDRA).**

I. INTRODUCTION

Computer Aided Diagnosis (CAD) Systems for melanoma detection usually follow three main steps: i) lesion segmentation; ii) feature extraction; and iii) lesion diagnosis [4], [6]. The vast majority of the systems extract features that describe four different aspects of the lesions: shape, symmetry, color distribution, and texture. These four types of features are inspired in the ABCD rule of dermoscopy, that takes into account asymmetry (A), border (B), number of colors (C), and differential structures (D) [12]. The extracted features are usually combined into a single feature vector and fed into a classification algorithm to predict the lesion label (*e.g.*, melanoma or benign). CAD systems either assume that all the features are relevant for the decision or they select a subset of them, using a feature selection method (e.g, [4]).

Previous works [3], [10], [9] tried to answer this question and provide a comparative study on the relevance of the different types of features. The interesting results led to the formulation of a new problem: which is the best way to combine the different features - early of late fusion? Most works adopt an early fusion strategy, where the features are combined into a single feature vector. However, this may not be the optimal strategy. In this work we investigate which is the best strategy to combine different features: early fusion (concatenates all features into a single feature vector) or late fusion (combines the outputs of different classifiers, each one trained using a different type of feature). Studies have shown that late fusion is preferable when one is working with different types of features [11], as is the case of dermoscopy image analysis. Furthermore, early fusion often leads to feature vectors of large dimension, which are undesirable. A detailed assessment of early fusion for several types of features was recently presented in [9]. However, the comparison between early and late fusion was never performed in dermoscopy image analysis. The comparison is based on two different datasets: PH2 [8] and EDRA [1]. The first was acquired at a single hospital while the later was acquired at three different hospitals, with different acquisition setups. This allows us to compare the performance of the methods for single and multi-source datasets.

*This work was funded by grant SFRH/BD/84658/2012 and by the FCT project FCT [UID/EEA/50009/2013]

[1]C. Barata and J.S. Marques are with Institute for Systems and Robotics, Instituto Superior Tecnico, Lisboa, Portugal ana.c.fidalgo.barata@ist.utl.pt

[2]Department of Computer Science, Louisiana State University, Shreveport LA, USA

II. SYSTEM OVERVIEW

Fig. 1 shows the block diagram of a CAD system with early fusion of image features. Lesion segmentation is not the emphasis of this paper. Therefore, we use manual segmentations to separate the lesions from the healthy skin. The feature extraction and classification blocks will be addressed in this section while the feature fusion step will be addressed in the next section.

A. Feature Extraction

Several features can be used to characterize skin lesions (see [6] for a survey of the different features used in dermoscopy image analysis). In this work we explore two groups of features: global features, which represent the whole lesion by single feature vector, or local features, based on the partitioning of the lesion into smaller regions, each of them represented by a feature vector.

1) Global Features: this is the most popular type of features in dermoscopy image analysis. Global features include shape and symmetry descriptors (e.g., area, perimeter, and circularity measure), color descriptors (e.g., color histograms or the mean color), and texture descriptors (e.g., gradient related histograms) of a lesion [6]. In this work we use color features (color histograms, in three different color spaces: HSV, L*a*b*, and Opponent) and texture features (gradient's amplitude and orientation histograms, and Gabor filters). All of these features have been used with success in previous works [4], [5], [3], [9]. Shape and symmetry features are not used because these descriptors cannot be computed for lesions that are not fully contained in the image. This could lead to a reduction in the number of available training and test images (especially melanomas), which is undesirable.

Dermatologists pay a special attention to the border of the lesions. In order to include this knowledge in the CAD system, color and texture features are separately extracted from two areas of the lesion, namely the border and the inner part. The division of the lesion into these two regions is performed by eroding the lesion segmentation mask with a disk of radius r, which was empirically set to be $\frac{1}{10}$ of

978-1-4244-9270-1/15/$31.00 ©2015 IEEE

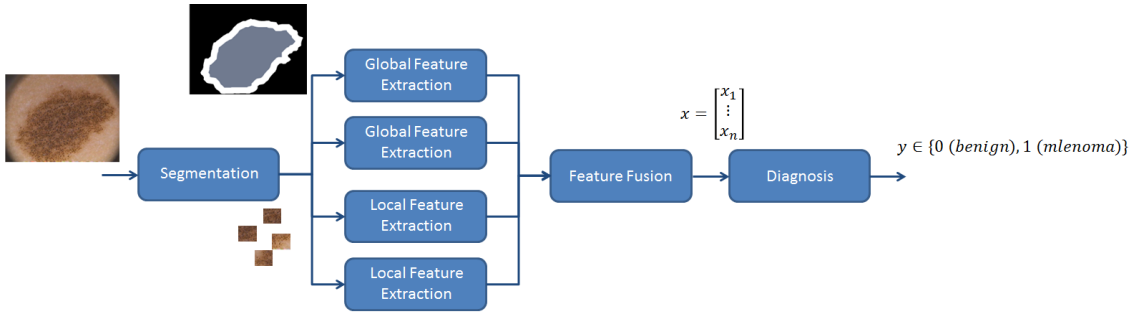

Fig. 1. Block Diagram of the CAD system. This scheme also depicts the early fusion strategy.

the lesion's minor axis. Figure 1 shows an example of this division. It was experimentally found in [3] that dividing the lesion into border and inner parts leads to a slight improvement of classification results using color and texture features.

2) Local features: these features allow us to efficiently characterize different regions of the lesion. A simple strategy to compute local features is the Bag-of-Features method, that works as follows. During the training phase the images are sampled into small patches, each of them characterized by a feature vector. In this work, we sample the images using a regular grid, obtaining image patches of size 40×40 pixels. Patches containing less than 50% lesion pixels are discarded. The color and texture descriptors used to characterize the patches are the same ones used as global features (see Table I for a summary of the used features).

The following step consists of clustering the feature vectors of the training set and obtaining a set of K centroids (called *visual words*) to represent the data. Traditionally the k-means algorithm is used in this step. Finally, the feature vectors are compared with the centroids and associated with the closest one. By counting the number of times each centroid occurs in the image, it is possible to build an histogram that characterizes the image. The clustering step occurs only once during the training phase, which means that the feature vectors of each image outside the training set are compared with the centroids obtained with the training set of images. For a detailed description of the role of local features please refer to [3], [2].

B. Lesion Classification

The last block of the system corresponds to the classification: melanoma or benign. During the training phase of the system, a classifier is learned using a set images previously diagnosed by a dermatologist. After this training step, the learned classifier can be applied to new images to predict the image class.

III. FEATURE FUSION STRATEGIES

The goal of this work is to compare possible schemes for feature fusion. This study is specially important since we can use different types of features to characterize skin lesions. Although most systems concatenate global features into a single feature vector, this may not be the correct option when one is working with global and local features at the same time. In the following sections we will describe the possible fusion and schemes and point out their strong and weak points.

TABLE I

FEATURES AND RESPECTIVE PARAMETER VALUES.

	Parameters	
Feature	Global	Local
Color Histograms - HSV (C1), L*a*b* (C2), and Opponent (C3) Spaces	32 bins per channel	16 bins per channel
Amplitude Histogram (T1)	16 bins	
Orientation Histogram (T2)	16 bins	
Gabor Filters (T3)	$N = 8$ orientations and 3 scales $\sigma_G \in \{2,4,8\}$	

A. Early Fusion

Fig. 1 shows the general scheme of a CAD system when early fusion is used. In this approach the different feature vectors are extracted and combined into a single representation. The easiest strategy consists of concatenating different feature vectors into a single one [7]. Then, the extracted vector is fed to a classifier to either learn the decision rule (training phase) or to predict the diagnosis (test phase).

An advantage of this strategy is that it performs learning and classification phases only once. However, the feature vector that results from the combination of features belongs to a high dimensional space. This might hamper the learning process (curse of dimensionality) and lead to the need of an additional feature selection step [7].

B. Late Fusion

Similarly to early fusion, this strategy also starts with the extraction of different features. However, in this case each type of feature is used to train a classifier. Then, the scores of the classifiers (assumed to be in the interval $[0,1]$) are combined in order to yield a final diagnosis. Different strategies can be applied to combine the scores of the trained classifiers [7]. In this work we compare two of them:

- **Majority Voting:** This is a simple strategy where the scores of the different classifiers are set to be either 0 (benign) or 1 (melanoma). A final decision is performed by counting the number of votes in each class, *i.e.*, the number of classifiers that gives 1 and those that give 0, and by selecting the class with the highest number of votes. This method is applied only if we have an odd number of classifiers
- **Supervised Learning:** A more elaborate approach consists of combining the scores of different classifiers

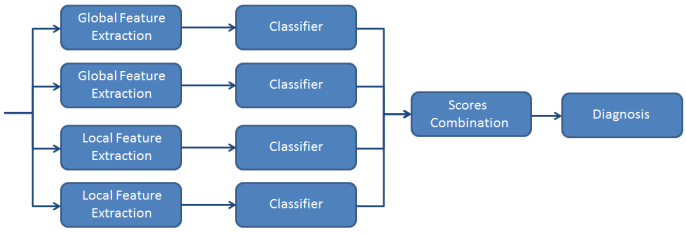

Fig. 2. General scheme for late fusion.

in a new feature vector that is fed to a new step of classification. This means that the training process comprises two stages of supervised learning. A first step in which set of classifiers are trained, each one using a different type of feature, and a second step in which the scores of the classifiers are used to train a final classifier to predict the diagnosis. During the test phase, the first set of classifiers produces the scores while the last classifier uses that info to obtain a final decision.

Fig. 2 shows the general scheme for late fusion. Notice that the "Feature Fusion" block and following steps in Fig. 1 are replaced by this new scheme. The main strength of this approach is the focus on the different performances of the used features. However, this technique is expensive in terms of learning effort, since it is necessary to train several classifiers [7].

IV. EXPERIMENTAL SETUP

The CAD systems were implemented and tested using two different datasets:

- **PH^2 dataset:** This is a publicly available dataset of 200 melanocytic lesions (40 melanomas), acquired at a single hospital [8].
- **EDRA dataset:** This dataset contains images collected at three different university hospitals: University Federico II of Naples (Italy), University of Graz (Austria) and University of Florence (Italy). Thus, EDRA is a multi-source dataset. We have selected 241 melanomas from this database (most of the available ones) and 241 benign lesions among Bluenevi, Clark nevi, Spitz nevi, Combined nevi and Dermal nevi.

For each of the datasets, we trained and tested different systems. We started by performing a separate evaluation of each type of feature (global or local) as well as the different descriptors (the three color histograms and the texture features). Then, the fusion of the different features was investigated. We developed several systems in order to consider all the possible feature combinations. Different classification algorithms (AdaBoost, SVM, kNN, and Random Forests) were tested in order to find the most suitable one. Random Forests yield the best results, thus, due to space constraints, we only discuss the results obtained with this algorithm. Each of the trained systems was optimized in order to give the best possible results. This means that for each case we varied a set of parameters, namely the number of trees $T \in \{1, 2, ..., 50\}$ and the number of centroids $K \in \{50, 100, ..., 300\}$, when using local features.

The performance of each system is evaluated using the following metrics: Sensitivity (SE) and Specificity (SP). SE corresponds to the percentage of melanomas that are correctly classified and SP is the percentage of correctly classified benign lesions. These metrics are computed using a nested-10 fold cross validation. The data was divided into 10 folds, with approximately the same number of melanomas and benign lesions. From these folds, 9 are kept for training and validation (selection of parameters) and the 10th fold is used for testing. The testing process is repeated ten times with a different fold, while the training- validation processes are performed nine times for each testing fold. Each time a different fold is kept out for validation. This process ensures that the choice of the best parameters is independent of the test set.

V. RESULTS

Table II shows the best results for each type of feature and descriptor. The considered features are the ones defined in Table I. These results are quite interesting. In both datasets the performance of color descriptors seems to be better when these are used as global features, while texture descriptors seem to perform better as local features.

TABLE II
LESION DIAGNOSIS RESULTS USING SINGLE FEATURES. BEST PERFORMANCE IN BOLD.

		Global		Local	
Dataset	**Feature**	SE	SP	SE	SP
PH^2	C1	92%	86%	92%	79%
	C2	92%	86%	94%	77%
	C3	**93%**	**84%**	92%	78%
	T1	84%	80%	87%	85%
	T2	61%	63%	90%	79%
	T3	90%	63%	**88%**	**88%**
EDRA	C1	77%	69%	72%	65%
	C2	**79%**	**69%**	71%	66%
	C3	73%	72%	68%	69%
	T1	73%	56%	**82%**	**56%**
	T2	74%	54%	78%	56%
	T3	74%	62%	79%	55%

The best early fusion results can be seen in Table III. This table shows the best combinations of global and local features, as well as the best final configuration. The results show that, as expected, it is possible to improve the performance of the system by combining more than

TABLE III
LESION DIAGNOSIS RESULTS USING EARLY FUSION. BEST PAIR CONFIGURATION/RESULTS - G STANDS FOR GLOBAL FEATURE AND L STANDS FOR LOCAL FEATURE.

		Results	
Dataset	**Combination**	SE	SP
PH^2	C2G + C3G + T2L +T3L	98%	87%
EDRA	C3G + T3G	80%	70%

TABLE IV

LESION DIAGNOSIS RESULTS USING LATE FUSION FOR MAJORITY VOTING AND SUPERVISED LEARNING. BEST PAIR CONFIGURATION/RESULTS - G STANDS FOR GLOBAL FEATURE AND L STANDS FOR LOCAL FEATURE.

			Results	
Dataset	**Combination**	**Strategy**	SE	SP
PH^2	C1G + T1L + T3L	Majority Voting	93%	95%
	C1G + C3G + T1L + T3L	**Supervised Learning**	**98%**	**90%**
EDRA	C1G + C2G + C3G + T2L + T3L	Majority Voting	83%	63%
	C1G + C2G + C3G + C2L + T2L + T3L	**Supervised Learning**	**83%**	**76%**

one type of feature. Interestingly, the best results for the EDRA dataset are achieved using only two features. We also noted that most of the setups that combined more than two features led to worse performances than the combination of only two features or even using one feature alone. This might be explained by the aforementioned problem of high dimensional feature vectors generated by combining features using early fusion.

The best late fusion results with majority voting and supervised learning can be seen in Table IV. Both late fusion approaches achieve good scores in the case of PH^2, while supervised learning outperforms majority voting in the case of the EDRA dataset. The best classification results are obtained with late fusion for both PH^2 (SE=98%, SP=90%) and EDRA (SE=83%, SP=76%) datasets. The overall results are also better than those obtained using early fusion. In this case, it was possible to combine more than two types of features and obtain better classification scores (*e.g.*, see the EDRA results). Since the late fusion approach consists of combining the outputs of different classifiers (recall Section III-B), it allows the use multiple descriptors without suffering from the curse of dimensionality that hampers early fusion. Another strength of late fusion (not investigated in this work) is the possibility to combine the outputs of different classifiers, *e.g.* combine the output of multiple SVM and Random Forests. This might improve the performance of a system, since we would be combining not only the strengths of different descriptors but also the strengths of different classifiers. The aforementioned strengths and the results suggest that late fusion might be the best strategy to be incorporated in a CAD system.

Examples of correctly classified lesions for both datasets can be seen in Figure 3. This images were correctly classified by the best late fusion systems highlighted in Table IV.

VI. CONCLUSIONS

The development of a CAD system for melanoma diagnosis requires the selection of appropriate features as well as the selection of the best strategy to combine them. In this work we have compared two different strategies for feature fusion: early and late fusion. The former is the one used in most CAD systems for melanoma diagnosis, while the later has never been used in this context. Our results have shown that late fusion method seems to be the best approach, with a SE = 98% and SP = 90% on the PH^2 and SE = 83% and SP = 76% on the EDRA datasets.

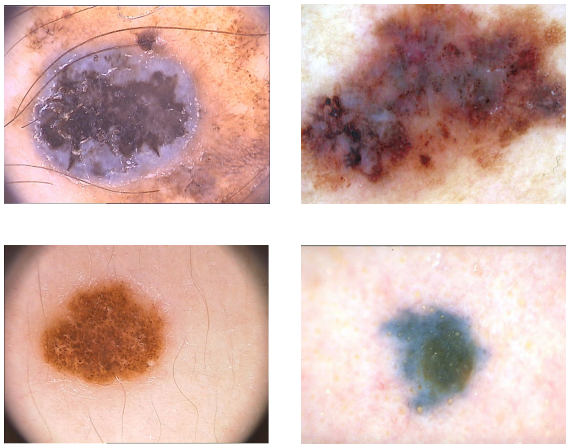

Fig. 3. Correctly classified melanomas (top) and benign (bottom) lesions. Examples for the PH^2 (left) and EDRA (right) datasets.

REFERENCES

[1] G. Argenziano and et al., *Interactive Atlas of Dermoscopy*. EDRA Medical Publishing & New Media, 2000.

[2] C. Barata and et al., "A bag-of-features approach for the classification of melanomas in dermoscopy images: The role of color and texture descriptors," in *Computer Vision Techniques for the Diagnosis of Skin Cancer*. Springer, 2014, pp. 49–69.

[3] ——, "Two systems for the detection of melanomas in dermoscopy images using texture and color features," *IEEE Systems Journal*, vol. 8, no. 3, pp. 965–979, 2014.

[4] M. Celebi and et al., "A methodological approach to the classification of dermoscopy images," *Computerized Medical Imaging and Graphics*, vol. 31, no. 6, pp. 362–373, 2007.

[5] H. Iyatomi and et al., "An improved internet-based melanoma screening system with dermatologist-like tumor area extraction algorithm," *Computerized Medical Imaging and Graphics*, vol. 32, no. 7, pp. 566–579, 2008.

[6] K. Korotkov and R. Garcia, "Computerized analysis of pigmented skin lesions: a review," *Artificial intelligence in medicine*, vol. 56, no. 2, pp. 69–90, 2012.

[7] L. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*. Wiley, 2004.

[8] T. Mendonça and et al., "Ph 2-a dermoscopic image database for research and benchmarking," in *35th EMBC*. IEEE, 2013, pp. 5437–5440.

[9] M. Rastgoo, R. Garcia, O. Morel, and F. Marzani, "Automatic differentiation of melanoma from dysplastic nevi," *Computerized Medical Imaging and Graphics*, vol. 43, pp. 44–52, 2015.

[10] M. Ruela and et al., "A system for the detection of melanomas in dermoscopy images using shape and symmetry features," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 2015.

[11] C. Snoek, M. Worring, and A. Smeulders, "Early versus late fusion in semantic video analysis," in *13th ACM on Multimedia*, 2005, pp. 399–402.

[12] W. Stolz, A. Riemann, and A. B. Cognetta, "ABCD rule of dermatoscopy: a new practical method for early recognition of malignant melanoma," *European Journal of Dermatology*, vol. 4, pp. 521–527, 1994.