Person Re-identification using Appearance: Final Report

Research Attachment

By:

Prerna Chikersal

Year 3, Computer Science,

Nanyang Technological University, Singapore

Supervised By:

Horesh Ben Shitrit

Prof. Pascal Fua

Computer Vision Laboratory,

EPFL, Switzerland

Abstract

The aim of this project is to apply a person re-identification algorithm which uses trivial appearance cues like color and texture, and achieve satisfactory performance without training, on various datasets being used by the lab. Firstly, a literature review was done and the tracking results were visualised. After this, we explored three methods - (i) dominant colors, (ii) color histograms and (iii) color invariants for person re-identification. We applied these methods on two sports datasets - a volleyball sequence and a soccer sequence, and on one pedestrian dataset - a video shot in the lab at EPFL. Based on the accuracies calculated, we concluded that the signatures used in the color invariants method produce the best results, since they are parts-based signatures or signatures which take spacial information into account. The dominant colors and color histograms methods do not work very well, since they are holistic approaches. Usually, the color histograms approach gives better results than the dominant colors approach, however, dominant colors can work better in a situation where the illumination changes are not much and consistent dominant colors can be obtained to describe a person or team. , like in the soccer dataset. Person re-identification based on appearance is a challenging problem, which works better if the clothes worn by the different people to be re-identified are very different in color and/or texture. In future, it would be interesting to see if superpixels can be used to divide the person into meaningful parts which can then be matched for person re-identification. Finally, the reports ends with acknowledgement and a synopsis of my personal experience.

1 Introduction

Person re-identification aims to recognise or re-identify a person who was previously seen in video(s) captured by one or more camera(s) with overlapping or disjoint field of views. This task has several applications in video surveillance, like tracking people in a crowd as they move across field of views of different cameras or in and out of the same camera's field of view. Re-identification allows us to reduce the possibility of identity switches by exploiting the appearance of a person to re-identify him/her again. In papers like [1, 2], person re-identification has been extended to the problem of tracking multiple players in sports videos whose paths may intersect and cause identity switches. This is a challenging problem due to occlusions and variations in illumination and pose caused by difference in viewing angle and lighting conditions as the person moves in a scene or from one field of view to another. [1, 3] use appearance cues like color histograms, jersey numbers and facial recognition to reduce identity switches.

2 Project Goal

The goal of this project is to apply a person re-identification algorithm which can exploit trivial appearance cues like color and texture to obtain satisfactory results without any training, on various datasets being used by the lab. We assume that jersey number recognition or facial recognition cannot be done, and instead, we explore other appearance cues which may be used for person re-identification.

In this report, we evaluate the effectiveness of using methods involving dominant colors matching, color histograms matching and matching using color invariants (parts-based shape descriptors, histogram over log-chromaticity color space, covariance descriptors and a combination of these) on two sports datasets - a volleyball game and a soccer game, and one pedestrian dataset - a video shot in the lab at EPFL. We primarily aim to classify players into teams for the sports datasets, while, we aim to carry out person re-identification for individuals for the pedestrian dataset.

3 Related Work

Researchers have attempted to solve the problem of person re-identification in various scenarios, such as tracking in pedestrian[4, 5, 6, 7] datasets or tracking sports players [1, 2], where the field of views of cameras may be overlapping [1] or non-overlapping [4, 5]. Moving cameras have also been used like in the ETHZ dataset used in [6].

In the past, to solve person re-identification, single-view and multi-view methods, as well as single-shot and multi-shot methods have been proposed. In single-view approaches, a person's appearance model is created based on information extracted from a single pose of the person, while in multi-view approaches, a person's appearance model is created based on information extracted from multiple poses. Futhermore, a single-shot approach is when a person's appearance model is created based on information from a single frame, while a multi-shot approach is when a person's appearance model is created based on information from multiple frames.

Doretto et al.[8] introduce the problem of person re-identification and briefly survey the different methods proposed to solve it. It talks about the major limitation of holistic appearance models based on histograms, which is that two people may have similar histogram-based signatures even though they are dressed differently. This can happen, if they have roughly the same amount of body surface covered with the same colors, regardless of the distribution of these colors. It then describes some methods which use appearance context modelling and parts-based modelling to overcome this problem. Parts-based approaches will be very useful for us, particularly when we want to create a unique signature for each player in a team.

Garcia et al. [4] extract information from multiple poses of a person along his/her trajectory to create his/her appearance model. Hence, this is a multi-view and multi-shot approach. They use tracked or known tracklets to find the orientation of the person's trajectory with respect to the camera. During tracking, the person's orientation and his/her feature vector at that time is stored. After tracking is lost, the distance between feature vectors extracted from different orientations in the unknown tracklet and the stored feature vectors from known tracklets that have a similar range of orientations is calculated. The lesser is the distance between the tracklets, the higher is the probability that they belong to the same person.

Kuo et al. [5] use multiple frames, but do not consider multiple poses of the person along with his/her trajectory. Hence, this is a single-view and multi-shot approach. Tracklets which do not intersect but exist in the same frames can never belong to the same person. Hence, features are extracted from these tracklets and taken as negative samples. Whereas, features extracted from two different responses of the same track are considered to be positive samples. Features from these negative and positive samples are used as input to Adaboost which carries out binary classification, which given 2 tracklets, determines if they belong to the same person or to different people.

Farenzena et al. [6] describe a parts-based approach in which they divide each person into parts using two lines of asymmetry, which separate the head, torso and legs of the person, and one line of symmetry, which divides these three parts into two subparts each. This paper also mentions the three ways in which we can perform feature matching (i) single-shot vs single-shot, that is each image represents a different individual in both, the gallery and the probe set, (ii) multiple-shot vs single-shot, that is each image in the probe set represents a different individual while each individual in the gallery in represented by multiple images, and (iii) multiple-shot vs multiple-shot, that is both gallery and the probe set contain signatures from multiple images.

Liu et al. [7] explore different sets of features, in an attempt to determine, what features are important in person re-identification. They use an unsupervised approach for learning a bottom-up feature importance, so features extracted from different individuals are weighted adaptively, driven by their unique and inherent appearance attributes. They conclude that instead of biasing all weights to features like color that are globally important to all individuals, it is better to selectively distribute some weights to features specific to certain appearance attributes of a person. For example, texture will be more important for a person wearing a textured shirt, while it will usually be irrelevant for a person wearing a plain, textureless shirt.

In [9], the authors segment a person's image into super pixels, from which they extract C-SIFT local features as visual words and build a TF-IDF vocabulary index tree to speed up people search. Finally, they adopt an image-retrieval way to implement person re-identification. This method requires the vocabulary tree to be built offline using training images. In our project, we wish to devise a completely online method for person re-identification which requires no training.

Kviatkovsky et al. [10] use an invariant colour signature to match people. They mark the observations from the upper/lower parts with red/blue markers and plot them onto a log-chromaticity histogram, and found that for each person, two main modes are generate, and the distribution structure for the same person is sufficiently preserved, while the distribution structure for different people is different. We will elaborate more on this method further on, in this report.

Most person re-identification methods like [2, 9] require some training. Training samples are collected online in [5]. While, methods like [4, 6] require no training. However, we can see that the results of methods that require no training are not very good. In this report, we will evaluate the performance of methods requiring no training.

4 Methods

Person re-identification methods aim to find an accurate way of (i) calculating the distance between persons and (ii) being able to tell which person is similar to an already observed person and which person is new or dissimilar to all persons observed before, based on the distances calculated. We want to be able to do this online, without the need for any training. We will be using a single frame to initialise all our methods, hence (ii) is not required.

4.1 Tracking multiple people and Background Subtraction

We use the algorithm described in [1] to track multiple people in multiple cameras. As input, we are given a Probabilistic Occupancy Map (POM) [11] containing probabilities of presence of people in each grid cell, which can be generated by any people detector. K-Shortest Paths (KSP) algorithm is used to find trajectories which may include identity switches, but we are able to find the grid cells in which we expect to find people at any given time. The tracking results are stored in a text file. Figure 1 shows the tracking results for the lab sequence videos. We use the background subtraction algorithm used in [11].

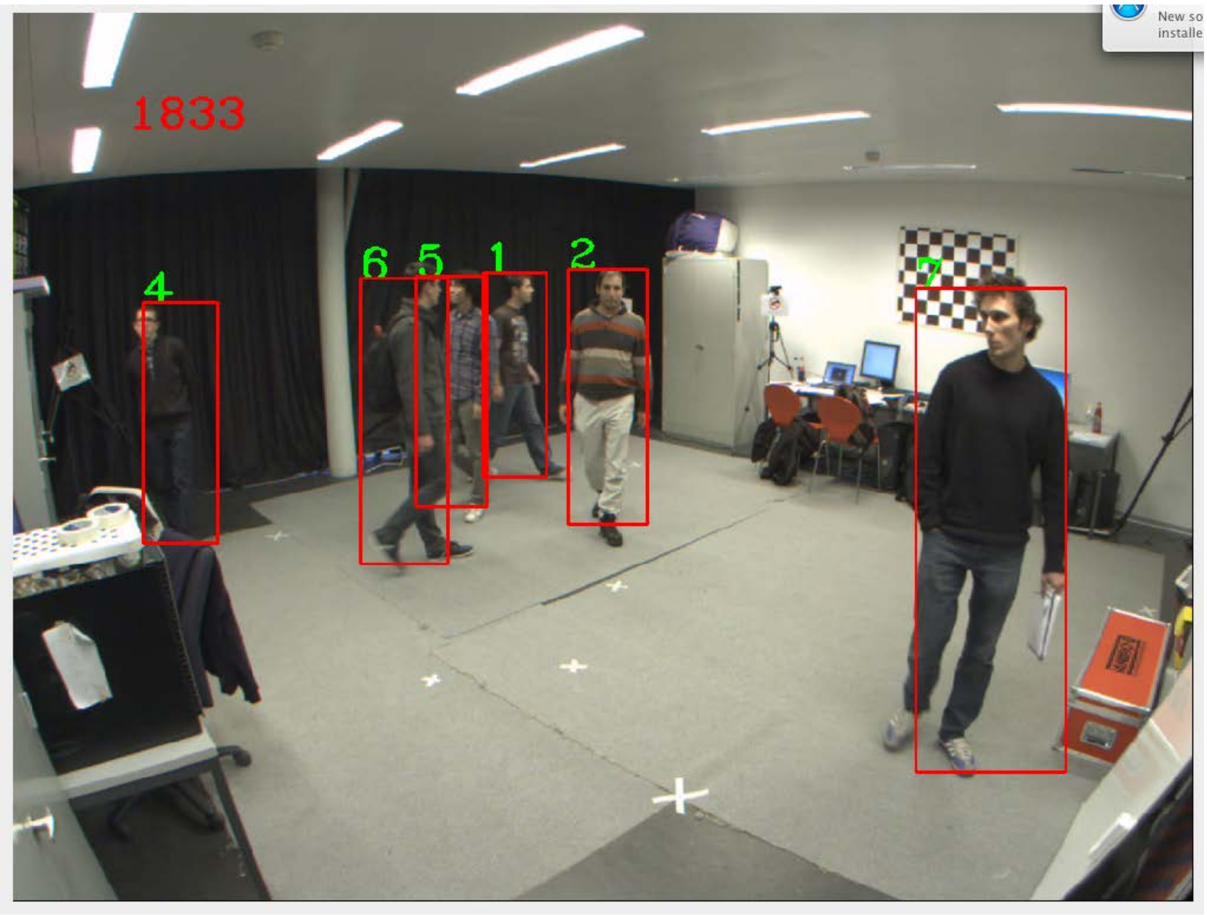

Figure 1: Tracking results from [1]

4.2 Dominant Colors

Dominant colors are the most occurring colors in an image. The primary dominant color is the first most occurring color, the secondary dominant color is the second most occurring color, while the tertiary dominant color is the third most dominant color, and so on.

We consider each pixel to be a point in 3D color space where the 3 axises represent red, green and blue colors respectively. We use k-means clustering based on euclidean distance to cluster these pixels into 4 clusters. The centroids of these 4 clusters are the 4 most dominant colors.

We use the dominant colors method to classify players into teams in the volleyball sequence. To calculate the distance between players in a multi-camera environment, we associate a description vector with the teams in each camera. Therefore, for 3 cameras, we have 3 description vectors which are of length n, assuming n is the number of teams. A frame is manually chosen and the dominant colors of players in each team in that frame is calculated and stored in a file. The file is loaded to initialise the description vectors.

Now, for each subsequent frame received at real-time, we first find the dominant colors for each player using k-means clustering, such that, we get 4 dominant colors representing each player queried. The number of colors in each team descriptor is a multiple of 4 and every set of

4 colors belongs to the same player in some frame. For each of the 4 colors of the player queried, we iterate through the team description vectors of each team and find the set of 4 colors in the team description vectors which is closest to the set of 4 colors in the player queried. The sum of the euclidean distances between these colors is recorded. The distance becomes the calculated distance between the queried player and the team in question. As a result, we get the distance of the player from each team and the player is assigned to the team which it has the least distance with. If players are occluded in certain views or if we do not have enough information to calculate the dominant colors of the player in some camera views, we only consider those views for which the dominant colors of the player can be successfully found.

In the single-shot approach, the colors of the player queried is compared only with the colors the teams description vector was initialised with. Whereas, in the multi-shot approach, every time a person is assigned to a team, his/her colors are also added to the teams description vector.

In the case, where the scene is not a closed system, that is, people can move in and out of the system such that the number of people vary and new people also enter the scene, we still have to manually describe each team which can be seen in the video sequence in the file which is used to initialise the team description vectors.

4.3 Color Histograms

Holistic color histograms are used to prevent identity switches in [1]. In this project, we also explore the effectiveness of color histograms in person re-identification.

A vector of color histograms for people is created and is initialised with the color histograms received in the first frame. Lets call this vector our "gallery". For every subsequent frame, the color histograms of all the players are calculated and compared to the histograms in the gallery. The distance between the queried histogram and each histogram in the gallery is calculated. If the least distance is smaller than a certain manually specified threshold, the queried histogram and the closest histogram in the gallery are considered to be of the same person. Else, the queried histogram is considered to be the histogram of a new person who has not been observed before. This threshold is chosen by hit and trial and can vary from dataset to dataset. In a closed system, we can initialise the person description vectors using one frame and we will not need to find a threshold, since the person queried is always considered to be the same as the person in the gallery whose histogram it is closest to.

4.4 Color Invariants

Kviatkovsky et al. [10] extracted invariant colour signatures from images of people for person re-identification. They devised a parts-based shape descriptor called PartsSC, which is based on the premise that if we mark the observations in log-chromaticity color space from the upper/lower parts of a person in red/blue, the distribution structure or shape obtained will be different for different people and similar for the same person. The four signatures used are:

1. *PartsSC:* A parts based shape descriptor is used to describe the distribution structure for each person. This descriptor encodes the relation between target colours and not between absolute colour values, and remains invariant to illumination changes, according to the diagonal model.

2. *Hist:* Signatures like Histogram over log-chromaticity colour space are used to capture Absolute colour values. The Hist signature is also parts-based, that is, the distance between the upper parts and lower parts of two people are calculated separately and then added together.

3. *Cov:* A covariance descriptor is used to capture the pixel's spacial information and texture cues missed by the parts-based shape descriptor. The descriptor consists of RGB color values of the pixel and the spacial location of the pixel.

4. *Comb:* A combination signature is also evaluated by combining the parts-based shape descriptor, the histogram over log-chromaticity colour space and the covariance descriptor.

We store thumbnails of each person in every frame, and consider the instance of a person in every frame to be another pose of that person, such that, the number of cameras or poses is equal to the number of frames. Now, using [10], we get four distance matrix, which gives us the distance between the instance of every person in each frame with all people in every other frame, for each of the four aforementioned signatures. To calculate the accuracy, we take the first frame as a reference and calculate the correct matches between people in the first frame and people in the second frame to the last frame. We repeat this using frames second frame, third frame,....., last frame as reference one by one, and then average the accuracies computed in each case. Finally, we get the average accuracy obtained when using the PartsSC, Hist, Cov and Comb signatures.

5 Results

In this section, we explain how the ground truth and accuracy percentage is calculated for each dataset used. Then, we present the accuracy obtained for all the three methods applied on each dataset.

5.1 Ground Truth and Accuracy

We use the following datasets - (i) Volleyball dataset, (ii) Soccer dataset and (ii) Lab sequence. For the volleyball and soccer datasets, we have the ground truth for team classification, that is, we know that which player belongs to which team in every frame. While, for the lab sequence we have annotated the location of each person in every 5^{th} frame manually.

Every time, a player is correctly classified, it counts as a true positive, and every time a player is wrongly classified, it counts as a false positive.The final accuracy of the method is given by: $(TruePositives \div (TruePositives + FalsePositives)) \times 100\%$

5.2 EPFL Volleyball Dataset

Volleyball dataset (figure 2) is a sports dataset, comprising of several players belonging to different teams. Players belonging to the same team and the referees wear the same uniform (except the libero players), due to which person re-identification using appearance becomes very difficult. Hence, our goal for this dataset would be to achieve acceptable results for team classification, which means, we wish to classify the players or referees wearing the same uniform together.

Volleyball dataset is a closed system, that is, no player moves in or out of the field of view of the camera. There are many occlusions, since the volleyball court is smaller than the huge field in case of soccer.

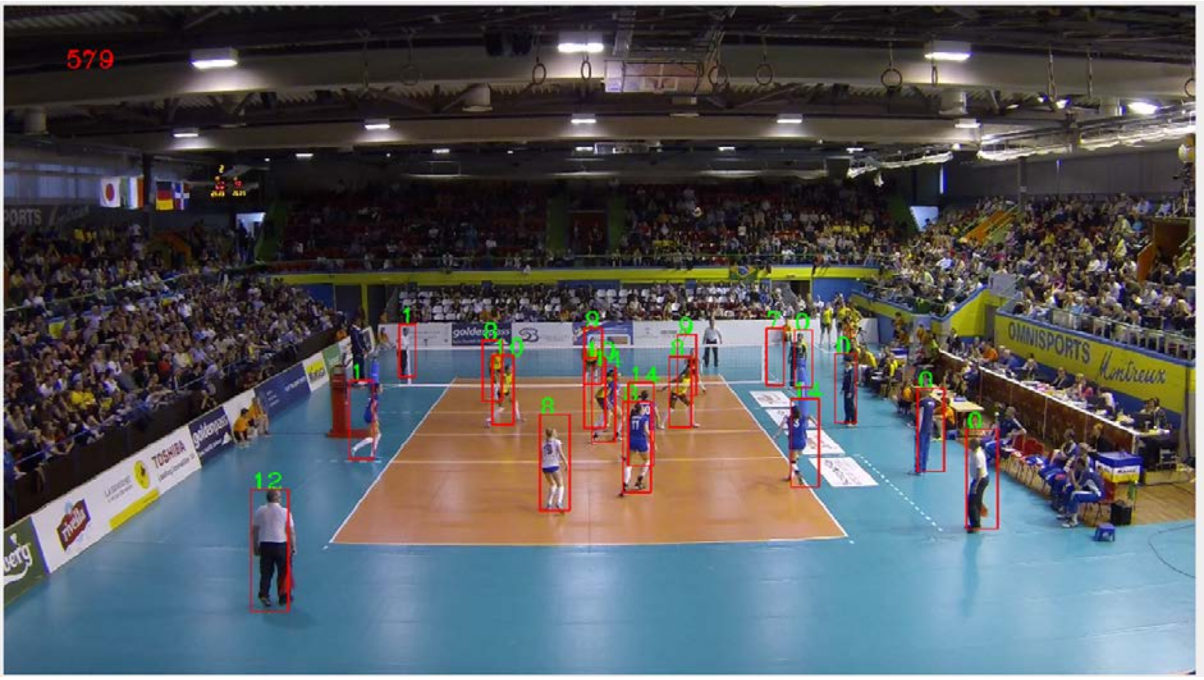

Figure 2: Snapshot from the EPFL volleyball dataset.

For this dataset, the results using dominant colors are obtained for the following cases: (i) Single-shot 3 videos, (ii) Multi-shot 3 videos, (iii) Single-shot 1 Video, and (iv) Multi-shot 1 Video. The results using color histograms and color invariants are obtained using the single-shot approach and only 1 video. Only team classification is performed using the dominant colors and color histograms, whereas, both team classification and person re-identification are attempted using the color invariants approach.

5.2.1 Results with Dominant Colors

The dominant colors approach was applied to the volleyball dataset. The maximum accuracy achieved over 1000 consecutive frames is 17.6965% which is quite low. Also, by observing the dominant colors extracted from the frames, we can say a single player might have a number of dominant colors in different frames. This is because of occlusions and the fact that in volleyball, the pose of the player varies greatly and illumination also changes as the player moves around in the scene.

Observations for the following are recorded and shown in figure 3:

1. *Single-shot 3 Videos:* means the team description vectors are initialised based on one frame or by manually describing the teams by specifying dominant colors for each team through a file. Once initialised, no changes are made to the description vectors. This is also a multi-view approach, since, 3 cameras with overlapping field of views are used.

2. *Multi-shot 3 Videos:* means the team description vectors are initialised based on one frame or by manually describing the teams by specifying dominant colors for each team through a file. After initialisation, every time, a player is classified into a particular team, his/her dominant colors are added to the description vector of that particular team. This is also a multi-view approach, since, 3 cameras with overlapping field of views are used.

3. *Single-shot 1 Video:* means the team description vectors are initialised based on one frame or by manually describing the teams by specifying dominant colors for each team through a file. Once initialised, no changes are made to the description vectors. This is a single-view approach, since, only one camera is considered for person re-identification.

4. *Multi-shot 1 Video:* means the team description vectors are initialised based on one frame or by manually describing the teams by specifying dominant colors for each team through a file. After initialisation, every time, a player is classified into a particular team, his/her dominant colors are added to the description vector of that particular team. This is a single-view approach, since, only one camera is considered for person re-identification.

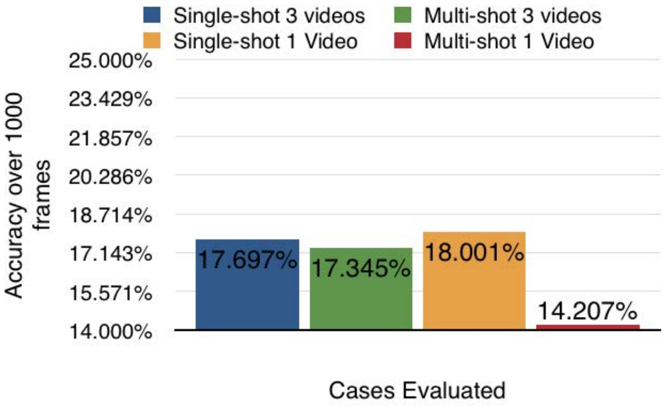

Figure 3: Accuracy of Dominant Colors method applied on volleyball dataset over 1000 consecutive frames. Four cases have been considered - Single-shot 3 videos, Multi-shot 3 videos, Single-shot 1 video and Multi-shot 1 video. Only team classification has been tried.

In order to compare with the more computationally intensive Color Invariants method, we also calculate the maximum accuracy over 10 consecutive frames using single-shot dominant colors for 1 video of the volleyball dataset, which is found to be 17.1123%.

5.2.2 Results with Color Histograms

Color histograms were used for person re-identification in the volleyball dataset as well. These experiments were only done with a single camera. For the volleyball dataset, we aim to classify the players into teams using color histograms. We only take the first 1000 consecutive frames and consider this to be a closed system where no body moves in or out of the field of view of the camera. The accuracy achieved by doing this is 41.2494%, which is better than the 17.6965% accuracy achieved using dominant colors.

In order to compare with the more computationally intensive Color Invariants method, we also calculate the maximum accuracy over 10 consecutive frames using color histograms for 1 video of the volleyball dataset, which comes to 32.8283%.

5.2.3 Results with Color Invariants for person re-identification

Since the color invariants method is very computationally intensive, we only evaluate this method over 10 consecutive frames. The colour invariants method computes four signatures as mentioned in 5.3.3.

For the volleyball dataset, our primary goal is to classify the players into teams. Person re-identification over many frames would be difficult in this scenario, since the players in the same team wear the same colored clothing. However, as we are using only 10 consecutive frames to evaluate this approach, we will try to do both team classification as well as person re-identification. Figure 4 shows us the results of this method on the volleyball dataset. Maximum accuracy for team classification is 90% and is obtained using the Hist signature, which uses histograms over log-chromaticity colour space to encode absolute color values. The Comb and PartsSC signatures are also not far behind with accuracies of 89.921% and 89.603% respectively for team classification. For person re-identification, the maximum accuracy is 89.365%, which is also obtained using the Hist signature.

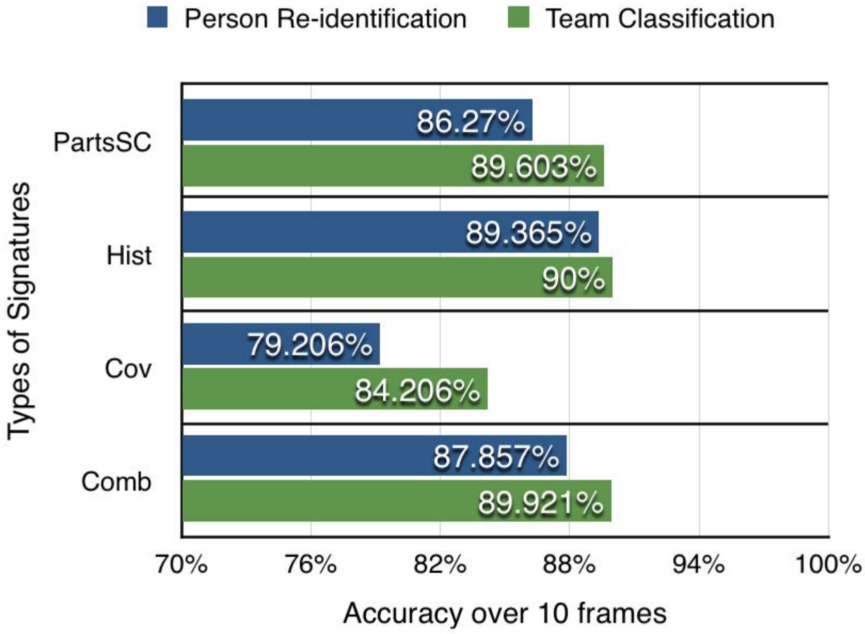

Figure 4: Accuracy of Color Invariants method applied on volleyball dataset over 10 consecutive frames. Four types of signatures have been used - PartsSC, Hist, Cov and Comb. Both team classification and person re-identification have been tried.

5.3 ISSIA Soccer Dataset

ISSIA Soccer dataset [12] (figure 5) is a sports dataset, comprising of several players belonging to different teams. Players belonging to the same team wear the same uniform, due to which person re-identification using appearance becomes very difficult. Hence, our goal for this dataset would be to achieve acceptable results for team classification, which means, we wish to classify the players or referees wearing the same uniform together.

Soccer dataset is an open system, that is, players move in or out of the field of view of the camera. The number of occlusions is lesser than the volleyball dataset and lab sequence, due to the huge size of the soccer field.

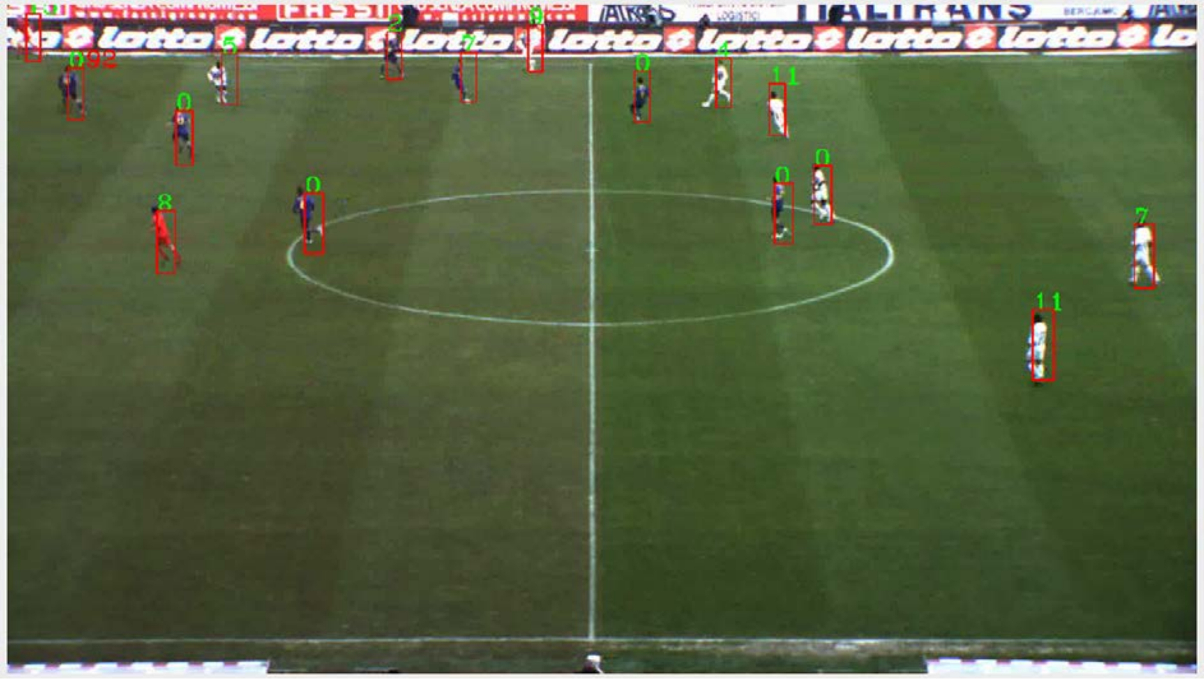

Figure 5: Snapshot from the ISSIA soccer dataset.

For this dataset, results for all three methods are calculated using the single-shot approach and only one video. Only team classification is performed using the dominant colors and color histograms, whereas, both team classification and person re-identification are attempted using the color invariants approach.

5.3.1 Results with Dominant Colors

The dominant colors approach was applied to the soccer dataset using the single-shot approach and only one video. The maximum accuracy achieved over 1000 consecutive frames is 31.62%, while, the maximum accuracy over 10 consecutive frames is 17.5097%.

5.3.2 Results with Color Histograms

Color histograms were used for person re-identification in the soccer dataset as well. These experiments were only done with a single camera. For the volleyball dataset, we aim to classify the players into teams using color histograms. Although, this is an open system, our primary goal is team classification and we assume that there is at least one player from every team in each frame. The maximum accuracy achieved over 1000 consecutive frames is 7.835%, while the maximum accuracy over 10 consecutive frames is 23.3333%.

5.3.3 Results with Color Invariants for person re-identification

The color invariants approach is applied to 10 consecutive frames taken from the lab sequence, in the same way as specified in 5.3.3. For the soccer dataset, our primary goal is to classify the players into teams. however, as we are using only 10 consecutive frames to evaluate this approach, we will try to do both team classification as well as person re-identification.

Figure 4 shows us the results of this method on the volleyball dataset. Maximum accuracy for team classification is 89.9921% and is obtained using the Hist signature, which uses histograms over log-chromaticity colour space to encode absolute color values. The Comb and PartsSC signatures are also not far behind with accuracies of 88.810% and 82.540% respectively for team classification. For person re-identification, the maximum accuracy is 71.667%, which is also obtained using the Hist signature, however the Comb signature is not far behind with an accuracy of 70.635% for person re-identification.

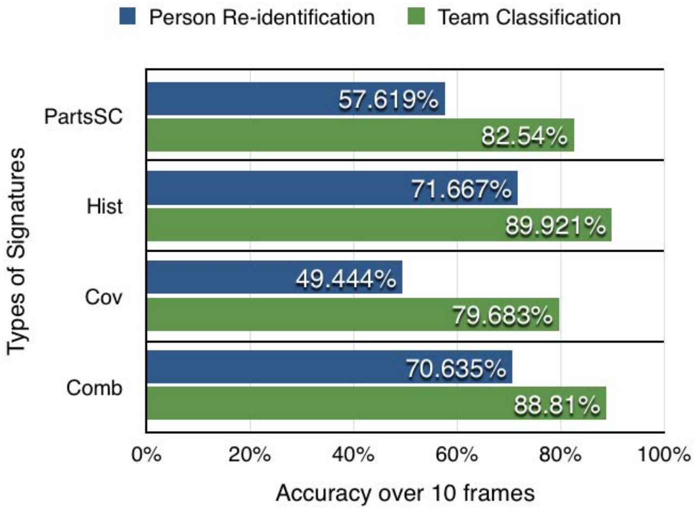

Figure 6: Accuracy of Color Invariants method applied on soccer dataset over 10 consecutive frames. Four types of signatures have been used - PartsSC, Hist, Cov and Comb. Both team classification and person re-identification have been tried.

5.4 EPFL Lab Sequence

Lab sequence (figure 7) is a video taken in a lab at EPFL. The sequence used for this experiment comprises of six individuals wearing different clothes. Our goal for this dataset is person re-identification, since team classification is not applicable as this is not a sports dataset. While the individuals wear different clothes, the color of the clothing is quite similar, which makes person re-identification using appearance a challenge.

The clip used is a closed system, that no body moves in or out of the field of view of the camera. There are many occlusions, since the space in the room is less and the people often cross paths.

For this dataset, results for all three methods are calculated using the single-shot approach and only one video. Team classification is not applicable, since this is a pedestrian dataset and

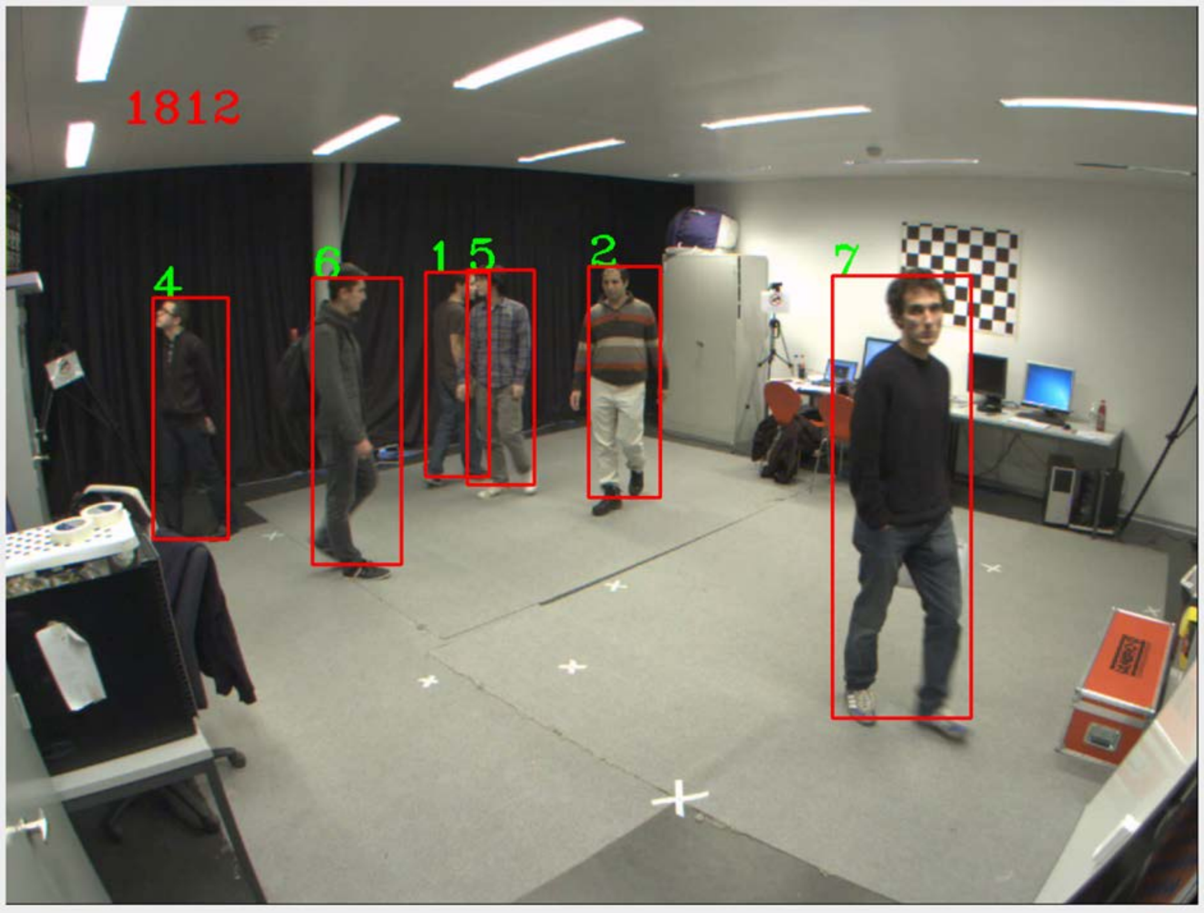

Figure 7: Snapshot from the EPFL lab sequence.

not a sports dataset. Hence, for all three methods, we aim to carry out person re-identification of individuals.

5.4.1 Results with Dominant Colors

The dominant colors approach was applied to a single view of the lab sequence over 1000 consecutive frames and 10 consecutive frames respectively. The maximum accuracy achieved for 1000 frames was 10.5712% and for 10 frames was 20.0%. The accuracy for this dataset is lower than the volleyball dataset because the main or dominant colors of the clothes of all the people in this dataset are more similar than those in the volleyball dataset.

5.4.2 Results with Color Histograms

Color histograms are also applied to the lab sequence. Since, there are no teams in the lab sequence, our aim is to identify individuals. This is similar to the team classification we did for the volleyball dataset, as the are classifying the histograms extracted for each player from each frame and classifying that into 6 persons, based on the distance between the histograms extracted from the frame and the histograms stored during initialisation. We can also consider this to be a closed system by taking only those frames in which the number of people remain constant. The accuracy achieved over 1000 consecutive frames is only 5.48193%.

Hence, classification based on color histograms work better for the volleyball dataset and barely work for the lab sequence.This is because, the colors of the clothing worn in the lab sequence are very similar to each other, while the colors of the players‘, referees‘ and coaches‘ uniforms are more different. These results are calculated by comparing our output with the available ground truth data.
In order to compare with the more computationally intensive Color Invariants method, we also calculate the maximum accuracy over 10 consecutive frames using color histograms for 1 video of the lab sequence, which comes to 11.6667%.

5.4.3 Results with Color Invariants for person re-identification

The color invariants approach is applied to 10 consecutive frames taken from the lab sequence, in the same way as specified in 5.3.3. However, since lab sequence is not a sports dataset and has six individuals, team classification is not applicable and we can only perform person re-identification. Figure 8 shows us the results of this method on the lab sequence. Maximum accuracy for person re-identification is 79.444% and is obtained using the Comb signature. PartsSC and Hist give us accuracies of 77.037% and 76.852% respectively, which are not too far behind from the maximum accuracy obtained by the Hist signature.

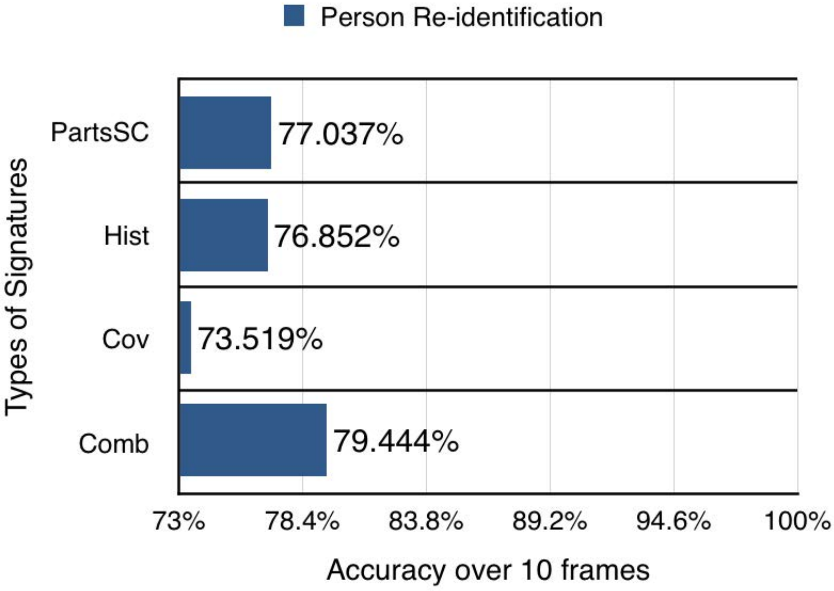

Figure 8: Accuracy of Color Invariants method applied on lab sequence over 10 consecutive frames. Four types of signatures have been used - PartsSC, Hist, Cov and Comb. Team classification is not applicable, as this is not a sports dataset, so only person re-identification has been tried.

6 Inference

In this section, we compare the various methods used and applied on different datasets. We try to determine which method would work best for different kinds of datasets, to achieve person re-identification and/or team classification.

6.1 Color Histograms vs Dominant Colors

The Color Histograms and Dominant Colors methods were applied on the volleyball dataset and soccer dataset for team classification and the lab sequence for person re-identification. Figure 9 shows the comparison between the accuracies obtained for 1 single-shot video of each dataset, over 1000 consecutive frames.

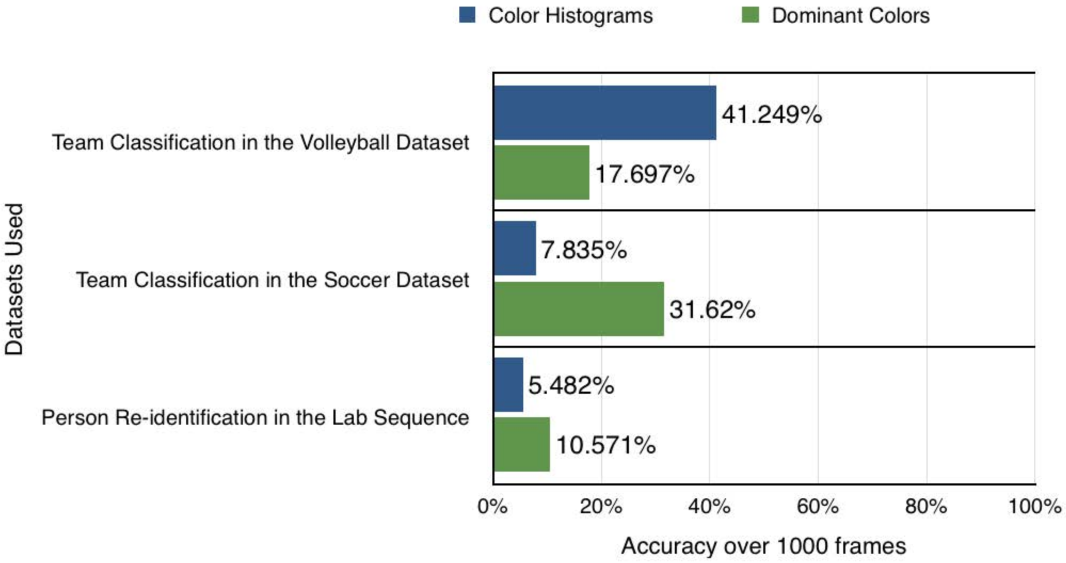

Figure 9: Comparison between the color histograms and dominant colors methods. The color histograms method does better than the dominant colors method in case of the volleyball dataset, while the dominant colors method does better than the color histograms in case of the soccer dataset. This could be because there are too many occlusions in the volleyball dataset or because the volleyball dataset is a closed system, but the soccer dataset is an open system, storing a color histogram as a template for each person in the soccer dataset, does not make sense, as the person may or may not be present in each frame. Also, for the soccer dataset, the dominant colors extracted from the clothes of every player are found to be much more consistent, than the volleyball dataset. For the lab sequence, both the datasets perform very badly and the difference between their accuracies is not very significant.

We can see that the color histograms method does better than the dominant colors method in case of the volleyball dataset, while the dominant colors method does better than the color histograms in case of the soccer dataset. This could be because there are too many occlusions in the volleyball dataset, due to which the dominant colors obtained for each person are not very accurate. Moreover, the volleyball dataset is a closed system, so storing color histograms as a template for each person, instead of dominant colors describing each team makes sense. However, since the soccer dataset is an open system, storing a color histogram as a template for each person, does not make sense, as the person may or may not be present in each frame. Figure 10 shows that color histograms perform slightly better over 10 consecutive frames. This could be because no player has moved in or out of the field in those 10 frames. Also, the quality of the volleyball dataset is better, due to which many shades of the same color can be seen across different frames, giving rise to many dominant colors for the same person or team. Whereas, the quality of the soccer dataset is poor and the dominant colors extracted from the clothes of every player are found to be much more consistent.

For the lab sequence, both the datasets perform very badly and the difference between their accuracies is not very significant.

6.2 Color Invariants vs Color Histograms vs Dominant Colors

The Color Invariants, Color Histograms and Dominant Colors methods were applied on the volleyball dataset and soccer dataset for team classification and the lab sequence for person re-identification. Color Invariants was also applied to the volleyball dataset and soccer dataset for team classification. Figure 10 shows the comparison between the accuracies obtained for 1 single-shot video of each dataset, over 10 consecutive frames.

For the volleyball dataset, the Hist signature used in the color invariants approach provides the best results for both team classification as well as person re-identification. PartsSC and the Comb signatures used in color invariants are also not very far behind, in both scenarios.

For the soccer dataset as well, the Hist signature used in the color invariants approach provides the best results for both team classification as well as person re-identification. For both team classification and person re-identification, the Comb signature is not very far behind and for team classification, the results obtained by the PartsSC signature are also very close.

Also, looking at the graph, it appears that person re-identification in the soccer dataset is harder than in the volleyball dataset, however, this may or may not be completely true, considering the fact that in the volleyball dataset, the players do not move much in 10 frames, while in the soccer dataset, the movement among players is much more. Therefore, it would be best to ignore the person re-identification results in case of the sports datasets, such that, team classification remains the main goal for the sports datasets.

For the lab sequence also, the Hist signature performs the best and PartsSC, Comb and Cov are all three not very far behind. However, the results obtained by the color histograms and dominant colors are very poor. If we compare all the datasets, we can see that the maximum accuracy obtained for the lab sequence is the worst (ignoring the person re-identification results of the soccer dataset). This is understandable, considering that the clothes worn by the people in the lab sequence do not differ much in color or appearance and therefore, person re-identification is comparatively harder.

7 Conclusion

In this section, we give a summary of what we concluded from this project, future work and my personal experience.

7.1 Summary

In this project, we evaluated different person re-identification algorithms for the datasets being used by the lab. We found that dominant colors can do better than color histograms when the illumination changes are not much and consistent dominant colors can be extracted for each person or team, however, both dominant colors as well as color histograms over the RGB color space do not produce admissible results on our datasets.

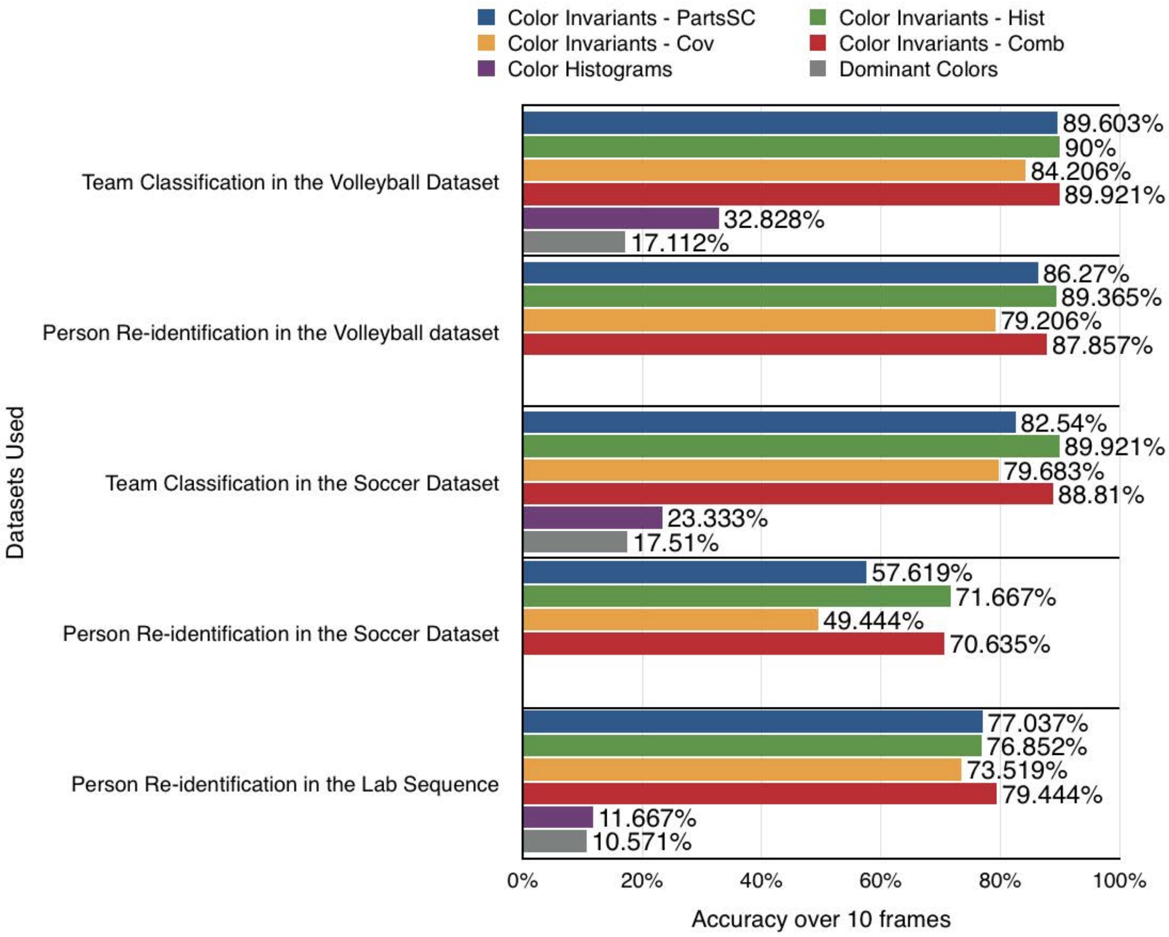

Figure 10: Comparison between the color invariants,color histograms and dominant colors methods. The Hist signature produces the best results in all 3 datasets. The PartsSC also performs very well and is almost as good as the Hist signature. The performance of the Cov signature is also quite good, except in case of person re-identification in the soccer dataset. Dominant colors and color histograms do not give acceptable results. All the four signatures in the Color Invariants method are either parts-based or take into account spacial information of pixels, while the dominant colors and the color histograms methods are holistic methods. From this, we can infer that parts-based approaches or approaches which take into account the spacial information are more accurate than holistic approaches.

Depending on the dataset, the Hist, PartsSC, Cov and Comb signatures used in Color Invariants usually return acceptable results. The Hist signature produces the best results in all 3 datasets and is created using histograms over the log-chromaticity color space. It is parts-based signature, that is, it divides the person into upper and lower parts. The PartsSC also performs very well and is almost as good as the Hist signature. This signature is a shape descriptor for colors and also divides the person into upper and lower parts. The performance of the Cov signature, which takes into account every pixels spacial information is also quite good. The dominant colors and color histograms approach were holistic approaches and did not give acceptable results. Hence, we can say that spacial or parts-based information is important in order to get good results for person re-identification.

We also inferred that since this approach completely relies on appearance cues, results obtained are better when the clothes worn by people in the video are quite different.

7.2 Challenges faced

Person re-identification using Appearance is a very broad topic and I was not very familiar with it. Hence, it took a long time to formulate and define the problem we intend to solve. Occlusions have not been ignored while evaluating any of the methods and are major problem for person re-identification, because when two people overlap or cross paths, there is no easy way to tell which part belongs to which person. Comparison between all 3 methods is done using only 10 consecutive frames, since the color invariants method is very computationally intensive. Not many occlusions occur in those 10 consecutive frames and so the color invariants method gives a high accuracy percentage. If we apply the color invariants method on a larger number of frames, the accuracy percentage might drop considerably. It should however still remain the best of all 3 methods.

The dominant colors and color histograms methods are both holistic methods. It would have been better to divide the person into parts and match each part separately. However, the problem here was trying to figure out how to divide a person into parts. The color invariants approach simply uses pre-defined regions of interest for the upper and lower body parts. We tried to divide the person into parts using superpixels, however, we soon realised that superpixels by themselves are meaningless and we will need to find a way to group superpixels together into meaningful parts. To do this, we tried using the human body parts detector in [13], but the detected parts were very inaccurate and so we decided to forego the use of superpixels or the human body parts detector.

While, the volleyball dataset is of better quality than the soccer dataset, many shades of each color can be seen in it, due to better clarity. This is a problem for the dominant colors method, since the dominant colors extracted for each person or team were very inconsistent for the volleyball dataset. The dominant colors extracted from the soccer dataset are much more consistent.

7.3 Future work

Presently, the color invariants approach is computationally intensive and not suitable for real-time applications. In future, it would be nice to modify this algorithm, such that, it can be used on videos in real-time.

Also, earlier, we had planned to use superpixels to divide each person into parts, however, superpixels by themselves cannot divide the person into meaningful parts for matching or comparison. In future, it would be interesting to see how we can use the results obtained by superpixel segmentation to divide each person into meaningful parts for matching. Human body part detectors may be used for this purpose, however, we did not find them to be very accurate.

7.4 Acknowledgement and My Personal Experience

I would like to thank my supervisor, Horesh Ben Shitrit for making this experience a fruitful learning experience for me. His guidance, insights and comments have indeed been very valuable! I would also like to thank Prof. Pascal Fua and my home University, Nanyang Technological University, Singapore for giving me the opportunity to do a research attachment at Computer Vision Laboratory (CVLab), EPFL.

This project was my first research project in Computer Vision. The literature review and methods experimented with, gave me a good understanding of the problem of person re-identification. I learnt how to formulate a research problem from scratch and how to compare or evaluate various methods, and derive useful conclusions from the results obtained. This experience will prove useful when I do my final year project and other projects at NTU. It will also help me decide the career path I would like to follow, after I graduate.

I had done a summer internship at Computer Graphics Laboratory (LGG), EPFL, before I started working at CVLab. So, I've been here for around 8 months now. Apart from being a great academic experience, working and staying in a Switzerland was an amazing experience in itself. Switzerland is one of the most beautiful countries I've ever been to. I also got a chance to travel to other neighbouring European countries and learn about different places, cultures and people, thus becoming more independent as an individual. I am glad that I decided to come here, for this experience has impacted me positively, in academics and otherwise.

References

[1] H. Ben Shitrit, J. Berclaz, F. Fleuret, and P. Fua, "Tracking multiple people under global appearance constraints," in *Computer Vision (ICCV), 2011 IEEE International Conference on*, pp. 137–144, IEEE, 2011.

[2] W.-L. Lu, J.-A. Ting, K. P. Murphy, and J. J. Little, "Identifying players in broadcast sports videos using conditional random fields," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 3249–3256, IEEE, 2011.

[3] H. BenShitrit, M. Zervos, P. Fua, F. Fleuret, *et al.*, "Facial descriptors for identity-preserving multiple people tracking," tech. rep., 2013.

[4] J. García, C. Kambhamettu, A. Gardel, I. Bravo, and J. L. Lázaro, "Features accumulation on a multiple view oriented model for people re-identification," in *Eurographics 2013 Workshop on 3D Object Retrieval*, pp. 105–108, The Eurographics Association, 2013.

[5] C.-H. Kuo, C. Huang, and R. Nevatia, "Multi-target tracking by on-line learned discriminative appearance models," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 685–692, IEEE, 2010.

[6] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 2360–2367, IEEE, 2010.

[7] C. Liu, S. Gong, C. C. Loy, and X. Lin, "Person re-identification: what features are important?," in *Computer Vision–ECCV 2012. Workshops and Demonstrations*, pp. 391–401, Springer, 2012.

[8] G. Doretto, T. Sebastian, P. Tu, and J. Rittscher, "Appearance-based person reidentification in camera networks: problem overview and current approaches," *Journal of Ambient Intelligence and Humanized Computing*, vol. 2, no. 2, pp. 127–151, 2011.

[9] C. Liu and Z. Zhao, "Person re-identification by local feature based on super pixel," in *Advances in Multimedia Modeling*, pp. 196–205, Springer, 2013.

[10] I. Kviatkovsky, A. Adam, and E. Rivlin, "Color invariants for person re-identification," 2012.

[11] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua, "Multicamera people tracking with a probabilistic occupancy map," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, no. 2, pp. 267–282, 2008.

[12] T. D'Orazio, M. Leo, N. Mosca, P. Spagnolo, and P. L. Mazzeo, "A semi-automatic system for ground truth generation of soccer video sequences," in *Advanced Video and Signal Based Surveillance, 2009. AVSS'09. Sixth IEEE International Conference on*, pp. 559–564, IEEE, 2009.

[13] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures-of-parts," 2012.