Skoltech
Skolkovo Institute of Science and Technology

Thesis Changes Log

Name of Candidate: Evgeniya Ustinova

PhD Program: Computational and Data Science and Engineering

Title of Thesis: Image-based human re-identification and recognition using deep learning methods

Supervisor: Prof. Victor Lempitsky

Chair of PhD defense Jury: Prof. Ivan Oseledets ***Email**: i.oseledets@skolkovotech.ru*

Date of Thesis Defense: 11 December 2019

The thesis document includes the following changes in answer to the external review process.

Dear Reviewers,
I would like to thank you for all the useful comments and thoughtful questions. I will be happy to continue the discussion during the thesis defense.
Please find the responses below.
Yours sincerely,
Evgeniya Ustinova

Reviewer: Ondra Chum

C1: It would be interesting to show, whether combining features from different regions (e.g., neighbouring) would be beneficial or not.
R: Combining features from different locations could indeed be an interesting direction. It is also an important question, how to choose appropriate sets of locations. In this work we, however, evaluate the simplest variant, namely, combining similarly located local features.

C2: Is it important to share parameters of the convolutions between the parts, or would part-specific convolutions improve the accuracy?
R: We used the state-of-the-art architecture from the paper [Yi et al., 2014a] as a baseline, and the particular version with shared first layer was implemented according to it. At the same time, the second layer use part-specific convolutions. I suggest that as the features learned by the first layer are rather low-level, making them part-specific should not improve the results.

C3: There are some problems with references that could be easily removed. For example, in the beginning of Chapter 3 and in Chapter 6, the formatting of references differs from the rest of the thesis (only the year is in the brackets).

R: The formatting in Chapters 3 and 6 has been fixed.

C4: Repeated incorrect reference to Section 1.3 as describing architectures, while the section introduces datasets. Three letter references that do not correspond to any paper, e.g., [Zhe] on page 49, [Kin] on page 52, etc. Reference to Chapter ?? on page 64.
R: I have fixed the references accordingly.

C5: In the caption of Fig. 6.2.: “last two rows” should be columns.
R: The caption of Fig. 6.2 has been fixed.

C6: It is a good practice to also report personal contributions for the publications listed as thesis related.
R: I have added personal contributions to the list of publications.

Reviewer: Prof. Ivan Oseledets
C1: Half of the thesis is related to the existing work and background, and the novel work starts from Chapter 3. It is not very clear, what is the challenge being solved by the author until that time. It would be nice to have main contributions of the thesis stated somewhere in the beginning.
R: Section 1.1 gives the introduction to the work, describes the context and the central tasks. Section 1.2 gives the background and motivation for each chapter. I have updated the beginning of Section 1.2 to make it more clear which tasks are being solved in the work. I have also updated the name of this section: “Motivation” → “Objectives and Motivation”.

C2: Personal contributions to the papers should be stated clearly.
R: Personal contributions have been added to the list of publications.

C3: The whole review part is nice, but it is not clear, if these results are needed for the results obtained in the thesis.
R: Indeed, the review section mostly gives context and shows the history of development of similarity learning and human recognition methods. However, Section 2.4 discusses several objective functions that are further used in the work.

C4: Each of the chapters covers an important problem, but can they be united by some common idea/method/task? This could be stated more explicitly.
R: Section 1.2 gives the picture of the general problem – building a deep learning framework for human recognition. However, I have renamed and updated it (see the response for the Comment 1).

Furthermore, Chapter 4 is based on the results of Chapter 3: the loss function introduced in Chapter 3 is used for all the experiments in Chapter 4 as it was demonstrated to show the best performance for person re-identification. The results of Chapter 5 were chronologically the earliest
among all the results presented in this work, therefore methods from Chapter 3 and Chapter 4 were not used there. Although the contributions of each of the chapters are independent, they are all parts of building a person re-identification pipeline and can

be applied simultaneously.
Chapter 6 considers domain adaptation for surveillance face
recognition and uses the method from Chapter 5 as one of the baselines.

I have updated section 1.5 accordingly to show the connection between the chapters (page 13).

Reviewer: Prof. Radu Timofte
C1: Each such chapter can be seen as a standalone problem formulation, solution description and evaluation. This also is my main criticism: the proposed contributions while addressing the same human face recognition problem are rarely studied together.
R: To address this issue, I have updated section 1.5 (page 13) by discussing the connection between the chapters. However this still does not change the fact that all the contributions have not been studied together. Also, please see the response to Comment 5 of Prof. Ivan Oseledets.

Reviewer: Prof. Andrzej Cichocki
C1: Although some new results have been submitted to arXiv 2018 and 2019, however, no any significant papers have been published or accepted in the last two years.
R: The results of Chapter 6 were submitted to journal "Machine Vision and Applications" (impact factor 1.788) more than a year ago, have been revised, and unfortunately are still under review.

C2: Also in Bibliography, I could not find any references to very recent related works published in 2019 and quite few from 2018.
R: I have added a short discussion of the following works [Saquib Sarfraz et al., 2018, Suh et al., 2018, Kalayeh et al., 2018] to Section 2.5 of related work (page 33).

C3: Since the area of research is extremely competitive and hot/ popular and develops very fast, so now there exist several competitive works showing much better results than those demonstrated in this thesis. For example, the methods that explicitly utilize pose prediction or segmentation and more modern general−purpose architectures improve the results [Saquib Sarfraz et al., 2018, Suh et al., 2018, Kalayeh et al., 2018]. Therefore, in my opinion, the results presented in Chapter 4 become rather more historical and academic than practically useful today.
R: The results of person re-identification have indeed improved marginally over two past years, due to the use of powerful general-purpose architectures. However, at the moment of submission of the results for publication, deep learning methods for person re-identification did not clearly outperform some traditional approaches.

Reviewer: Prof. Evgeny Burnaev
C1: Does the accuracy of the histogram estimate influence the proposed loss for feature learning; if yes, then how?
R: We have conducted several experiments where different sizes of bins (from 0.001 to 0.04) were used to estimate the similarity histograms. The results are shown in Figure 3.2a: as it was expected, the quality is slightly worse for larger sizes of bins.

C2: should we somehow take into account confidence intervals, which can be calculated for the histogram estimate?
R: This would correspond to ignoring the hardest examples (negative pairs with high similarity and positive values with low similarity). Such techniques are indeed used for similarity learning, e.g., “semi-hard” mining for learning with triplet loss (2.11). Indeed, this is a promising direction for the future work.

C3: What if we use standard distances between distributions such as KL distance and maximize it during feature learning instead of the proposed histogram loss?
R: We tried K-L divergence for toy data, and it appeared not to perform well in the case when the initial distributions similarity values are highly overlapped. The reason is that it does not imply which distribution should be “on the left” and which should be “on the right”. That is why using K-L divergence may not lead to repelling signal for negative pairs and pulling signal for positive pairs.
I have added a short discussion on whether it is possible to use standard distances between distributions, like Kullback-Leibler divergence, instead of the suggested method (page 44).

C4: It could be good to discuss intuition behind the approach to fine-grained recognition.
R: The main intuition in the original work on Bilinear networks [Tsung-Yu Lin and Maji, 2015] was to use feature multiplication followed by global pooling to 1) factorize features to model “what” and “where” concepts, 2) ret rid of the problem of localization of important details. Person re-identification imply less pose variation than data used in the [Tsung-Yu Lin and Maji, 2015], so the idea of the Chapter 4 was that person re-identification performance could also be improved by Bilinear networks, but less radical pooling could be more appropriate. However, I should note, that state-of-the-art architecture used for person re-identification in Chapter 4 is shallower than that used in [Tsung-Yu Lin and Maji, 2015] and this could influence the results and conclusions.

C5: Why only two streams have been used for fine-grained recognition?
R: One problem is that using more streams would dramatically increase the number of parameters as the Bilinear layer outputs the quadratic number of features. Moreover, for additional experiments (not shown in the thesis) we have observed that for person re-identification, using output of only one stream for both inputs of bilinear layer results in similar performance as using two streams.