

## Personalized Recommendations in Police Photo Lineup Assembling Task

Ladislav Peška

Department of Software Engineering  
Faculty of Mathematics and Physics, Charles University, Prague  
Czech Republic  
peska@ksi.mff.cuni.cz

Hana Trojanová

Department of Psychology  
Faculty of Arts, Charles University, Prague  
Czech Republic  
trojhanka@gmail.com

**Abstract.** In this paper, we aim to present a novel application domain for recommender systems: police photo lineups. Photo lineups play a significant role in the eyewitness identification prosecution and subsequent conviction of suspects. Unfortunately, there are many cases where lineups have led to the conviction of an innocent persons. One of the key factors contributing to the incorrect identification is unfairly assembled (biased) lineups, i.e. that the suspect differs significantly from all other candidates. Although the process of assembling fair lineup is both highly important and time-consuming, only a handful of tools are available to simplify the task.

We describe our work towards using recommender systems for the photo lineup assembling task. Initially, two non-personalized recommending methods were evaluated: one based on the visual descriptors of persons and the other their content-based attributes. Next, some personalized hybrid techniques combining both methods based on the feedback from forensic technicians were evaluated. Some of the personalized techniques significantly improved the results of both non-personalized techniques w.r.t. *nDCG* and *recall@top-k*.

### 1 Introduction

Evidence from eyewitnesses often plays a significant role in criminal proceedings. A very important part is the lineup, i.e., eyewitness identification of the perpetrator. Lineups may lead to the prosecution and subsequent conviction of the perpetrator. Yet there are cases where lineups can played a role in the conviction of an innocent suspect. This forensic method consists of the recognition of persons or things and thus is linked with a wide range of psychological processes such as perception, memory, and decision making. Those processes can be influenced by the lineup itself. In order to prevent witnesses from making incorrect identifications, the lineup assembling task is among the top research topics of the psychology of eyewitness identification [1, 4, 6, 9, 10].

The sources of error in eyewitness identifications are numerous. Some variables affecting error probability are on the side of the witness (e.g., level of attention, age or ethnicity) and the event (e.g., distance, lighting, time of the day) and in general cannot be controlled [6, 9]. Controllable variables include the method of questioning, identification procedure, interaction with investigators, and similar [9, 10].

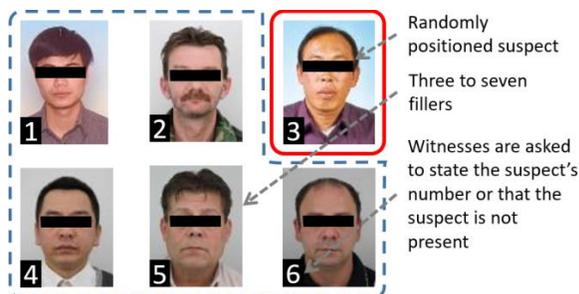
One of the principal recommendations for inhibiting errors in identification is to assemble lineups according to the lineup fairness principle [1, 5]. Lineup fairness is usually assessed on the basis of data obtained from "mock witnesses" - people who have not seen the offender, but received a short description of him/her. Lineup fairness measures a bias against the suspect and defining the

assembled lineup as fair if mock witnesses are unable to identify a suspect based only on a brief textual description. See Figure 1 for an example of a highly biased lineup.

Assembling photo lineups, i.e., finding candidates for filling the lineup for a particular suspect, according to the lineup fairness principle is a challenging and time-consuming task involving the exploration of large datasets of candidates. In the recent years, some research projects [4, 11] as well as commerce activities, e.g., *elineup.org*, aimed to simplify the process of eyewitness identifications. However, they mostly focused on the lineup administration and do not support intelligent lineup assembling.

From the point of view of recommender systems, lineup assembling is quite specific task for several reasons. Users of the system are respected experts, who assemble lineups regularly, although, usually, not on a daily bases. Therefore, we can expect a steady flow of feedback from long-term users. Also, each lineup assembling task is highly unique, i.e., the same suspect hardly ever appears in multiple lineups. Thus, some popular approaches incorporating collaborative filtering [2] or "the wisdom of the crowd" cannot be applied in this scenario. Last, but not least, the relevance judgement is highly based on the visual appearance and/or similarity of the suspect and lineup candidate.

In this paper, we describe our work in progress towards designing recommender systems aiding user to assemble fair lineups. In our previous work, we evaluated two non-personalized, item-based recommending strategies [8]. Based on the initial evaluation of non-personalized methods, we propose a content-based personalized approach combining both non-personalized techniques, aiming to re-



**Figure 1:** Example of an extremely biased lineup. Lineup usually consists of four to eight persons and witness is instructed that suspect may or may not be among them. However in this case, suspect can be easily identified even by a mock witness knowing only a short description such as, "Vietnamese male, 50-70 years old."

rank the list of proposed candidates according to the long-term preferences of the user.

More specifically, main contributions of this paper are:

- Proposed and evaluated hybrid personalized recommendation method.
- Dataset of assembled lineups with both positive and negative training examples.

To the best of our knowledge, our work is the first application of recommender systems principles on the lineup assembling task.

## 2 Item-based Recommendations

### 2.1 Dataset of Lineup Candidates

Although there are several commercial lineup databases<sup>1</sup>, we need to approach carefully while applying such datasets due to the problem of localization. Not only are the racial groups highly different e.g., in North America (where the datasets are mostly based) and Central Europe, but other aspects such as common clothing patterns, haircuts or make up trends vary greatly in different countries and continents. Underlined datasets should follow the same localization as the suspect in order to inhibit the bias of detecting strangers or having the incorrect ethnicity in a lineup. We evaluated the proposed methods in the context of the Czech Republic. Although the majority of the population is Caucasian, mostly of Czech, Slovak, Polish and German nationality, there are large Vietnamese and Romany minorities which make lineup assembling more challenging. We collected the dataset of candidate persons from the *wanted and missing persons* application<sup>2</sup> of the Police of the Czech Republic. In total, we collected data about 4,423 missing or wanted males. All records contained a photo, nationality, age and appearance characteristics such as: (facial) hair color and style, eye color, figure shape, tattoos and more. More information about the dataset may be found in [8].

### 2.2 Item-Based Recommending Strategies for Lineup Assembling

In our previous work [8], we proposed two non-personalized recommending strategies, where the list of proposed candidates is based on the similarity between the suspect and lineup candidates. We use the underlined assumption that the lineup fairness can be approximated through the similarity of the suspect and fillers, i.e. by filling lineups with candidates similar to the suspect, we ensure that lineups remain unbiased.

*Content-based Recommendation Strategy (CB-RS)* leverages the collected content-based attributes of candidates. We employed the Vector Space Model [3] with

binarized features, TF-IDF weighting and cosine similarity. *CB-RS* strategy was intended to be closely similar to the attribute-based searching, which is commonly available in lineup assembling tools.

*Recommendation Based on visual features (Visual-RS)* leverages the similarity of visual descriptors received from a pre-trained CNN (VGG network for facial recognition problems, VGG-Face [7], in our case). More information is available in the previous work [8].

### 2.3 Evaluation of Item-Based Recommenders

To make this paper self-contained, let us briefly describe the results of non-personalized recommendation strategies.

The evaluation was based on a user study of domain experts, i.e., forensic technicians, whose task was to select best lineup candidates out of the ones recommended by both techniques. More specifically, 30 persons were selected from the dataset to play the role of suspects. For each suspect, both non-personalized recommendation strategies proposed top-20 candidates that were merged into a single list<sup>3</sup> and displayed together with the suspect to the domain experts. Domain experts selects the most suitable candidates; these were considered as positively preferred. Participants were instructed to maintain lineup fairness principles, they were allowed to produce incomplete lineups if no more suitable candidates were available, or select more candidates if they were equally eligible.

The evaluation was performed by seven forensic technicians from the Czech Republic, with 202 assembled lineups and 800 selected candidates in total. Table 1 illustrates overall results of the user study. One can observe that although *Visual-RS* clearly outperformed *CB-RS*, also the candidates recommended by *CB-RS* were selected quite often. Together with the surprisingly low size of the intersection (1.83%) between the lists of recommended candidates and relatively high level of disagreement among participants on the selected candidates, the results indicate that some merged, personalized strategy is plausible. Furthermore, as the mean rank of selected candidates was

**Table 1: Evaluation results depicting the volume of selected candidates, the differences in volumes of selected candidates (p-value of paired t-test), the level of agreement among participants (Krippendorff's alpha) and the average rank of the selected candidates. Note that candidates proposed by both strategies were excluded from results.**

	Selected candidates	P-value	Level of agreement	Average rank
<i>Visual-RS</i>	466 / 58%	1.2e-8	0.178	8.2
<i>CB-RS</i>	298 / 37%		0.138	8.9

decision whether the next list item will be filled by *CB-RS* or *Visual-RS* method.

<sup>1</sup> e.g., <http://elineup.org>

<sup>2</sup> [aplikace.policie.cz/patrani-osoby/Vyhledavani.aspx](http://aplikace.policie.cz/patrani-osoby/Vyhledavani.aspx)

<sup>3</sup> The ordering of candidates proposed by each method was maintained, i.e., the randomness was applied on the

relatively high for both methods (8, resp. 9 out of 20), there is a room for some re-ranking approach.

### 3 Personalized Recommendations

Based on the evaluation of non-personalized, item-based recommending techniques, we hypothesized that the proposed recommendations can be further improved by employing some content-based personalized techniques. We approach this task through state-of-the-art machine learning methods as follows.

Suppose that for arbitrary user  $u$ , his/her previous interactions with the system are in the form of triples  $F_u: \{r_u(i, j)\}$ , where  $i$  is the suspect of some previously created lineup,  $j$  is a recommended candidate and  $r_u = 1$  if  $j$  was selected to the lineup and  $r_u = 0$  otherwise. Furthermore, both  $i$  and  $j$  can be represented by three sets of attributes:

- $A^{cb}$  are TF-IDF values of content-based attributes of each object.
- $A^{vis}$  represents the visual descriptor based on the VGG-Face network.
- The union of both sets:  $A^{cb} \cup A^{vis}$

Suppose that equations below represents scoring functions of the non-personalized recommending strategies.

$$s_{cb}(i, j) = \frac{1}{1 + \sum_{a \in A^{cb}} |a_i - a_j|} \quad s_{vis}(i, j) = \frac{1}{1 + \sum_{a \in A^{vis}} |a_i - a_j|}$$

Now, let us define a personalized classification / regression task<sup>4</sup> with the train set examples constructed as follows. For each  $f \in F_u$ , the output variable  $y = r$  and the list of dependent variables  $\mathbf{x}_A$  are constructed as a subtraction of suspect's and candidate's attributes for a set of attributes  $A$ :  $\forall a \in A: x_a := |a_i - a_j|$ .

Given an arbitrary classification method  $M$ , the model of user preferences  $m_{u,A}$  is trained by applying method  $M$  on the per-user train set  $\{(\mathbf{x}_A, y)\}$ . When the user starts a new lineup task with some new suspect  $\bar{i}$ , the lineup candidates are ranked according to their probability to be selected in the lineup:

$$r_j := P(r_u(\bar{i}, j) = 1 | m_{u,A}).$$

We would like to note that such recommendation scenario is quite challenging as we do not have any feedback from the current lineup and need to rely solely on the long-term user preferences (note the relation to the page zero problem or homepage recommendation problem). On the other hand, quite complex learning methods can be used, because the time-span between two consecutive lineup assembling performed by the same forensic technician tends to be rather large.

Following preference learning methods were evaluated<sup>5</sup>:

- Non-personalized similarity based on the  $L_1$  distances (baseline)
- Linear regression (denoted as LM in the evaluation)
- Lasso regression (Lasso)
- Decision tree (Dec. tree)
- Gradient boosted tree (GBT)

As the initial evaluation of the proposed method was only partially successful (machine learning methods were able significantly improve the baseline only in the case of  $A^{cb}$  attribute set), we further proposed a hybrid approach integrating two components:

- Predictions of a selected machine learning method on  $A^{cb}$  attribute set.
- Predictions based on a non-personalized  $L_1$  distance metric applied on  $A^{vis}$  attribute set.

Both prediction techniques are aggregated via probabilistic sum, i.e.,  $r_j := r_j^{cb} + r_j^{vis} - r_j^{cb} \times r_j^{vis}$ . This approach is denoted as *hybrid* in the evaluation.

### 3.2 Evaluation of Personalized Recommendations

The main goal of the personalized recommendations evaluation is to clarify, whether the long-term user preferences, i.e., collected during some previous lineups assembling, can be utilized to improve the list of recommended candidates for the current lineup.

In order to confirm this hypothesis, we performed an off-line evaluation on the dataset of assembled lineups collected during the evaluation of item-based recommendations. The resulting dataset contained in total 7659 records (800 positive and 6859 negative), i.e., in average 1094 records per user. Proposed methods were evaluated based on the 10-fold cross-validation protocol applied on the lineups. Hyperparameters of the methods were learned via grid-search on an internal leave-one-lineup-out protocol.

For each tested lineup, each recommending method re-ranks objects originally displayed to the forensic technicians according to the computed relevance  $r_j$  (selected candidates should appear on top of the list). We measure normalized discounted cumulative gain (nDCG), recall at top-10 and recall at top-5 (rec@10, rec@5 resp.) of the list and report on the average results for all evaluated users and lineups.

Table 2 depicts results of the off-line evaluation. We can observe that both linear model and gradient boosted trees improved over the baseline method in case of the  $A^{cb}$  attributes set. Therefore, we evaluated the hybrid approach with both methods. Both hybrid methods outperformed the best baselines w.r.t. nDCG and rec@5 metrics, while GBT hybrid provides the best performance w.r.t. all evaluated metrics.

<sup>4</sup> Please note that although the classification is a natural choice due to the binary output variable, the final output of the method should be ranking of candidates. Thus, we also evaluate several regression-based machine learning methods

and in case of classification method, we use positive class probability score as ranking.

<sup>5</sup> We use the methods' implementation from sci-kit package, <http://scikit-learn.org>.

**Table 2: Results of the personalized recommendation methods. Note that  $A^{vis}$  based machine learning approaches did not improve the baseline and were omitted for the sake of space.**

Method	Attributes	nDCG	rec@10	rec@5
Baseline	$A^{cb}$	0.4088	0.1796	0.0805
Baseline	$A^{vis}$	0.4990	0.3837	0.1725
Baseline	$A^{cb} \cup A^{vis}$	0.4201	0.2432	0.1090
LM	$A^{cb}$	0.4605	0.2949	0.1413
Lasso	$A^{cb}$	0.3816	0.1255	0.0484
Dec. tree	$A^{cb}$	0.3842	0.0871	0.0611
GBT	$A^{cb}$	0.4563	0.2728	0.1451
LM hybrid	$A^{cb} \cup A^{vis}$	0.4995	0.3693	0.2003
GBT hybrid	$A^{cb} \cup A^{vis}$	<b>0.5205</b>	<b>0.3843</b>	<b>0.2042</b>

#### 4 Conclusions

The main aim of this work in progress was to analyze the applicability of recommender systems principles in the problem of photo lineup assembling. Although the photo lineup assembling task is both important and time-consuming task, state-of-the-art tools do not provide intelligent search API beyond simple attribute search and to the best of our knowledge, apart from our work, there are no papers utilizing recommending principles in the lineup assembling task.

After the initial evaluation of item-based recommending algorithms, we proposed several variants of content-based personalized recommending algorithms utilizing long term preferences of the user. The off-line evaluation confirmed that long-term preferences can be used to improve the final ranking of candidates, however, only in case of content-based attributes.

Proposed approaches remained ineffective in the case of visual descriptors, so one direction of our future work is to further analyze this problem and providing solutions suitable also for visual descriptors. Siamese networks merging both content-based and visual descriptors seems particularly suitable for the task. Another option is to use visual descriptors as a base for short-term user preferences, i.e., the ones expressed in the current lineup and refine the recommended objects based on the already selected candidates.

Textual description of the suspect also plays an important role in the lineup assembling, as forensic technicians often tries to select candidates that match mentioned, highly specific, features, e.g., scars, skin defects, specific haircut etc. Another direction of our future work would aim to incorporate searching for these specific features in a “guided recommendation” API. Selecting specific regions of interest within the suspect’s photo seems to be a suitable initial strategy.

Finally, the long term goal of our work is to move from the recommendation of candidates to the recommendation of assembled lineups and to provide a ready-to-use software for forensic technicians.

#### Acknowledgments

This work was supported by the Czech grants GAUK-232217 and Q48.

#### References

- [1] Brigham, J.C. 1999. Applied issues in the construction and expert assessment of photo lineups. *Applied Cognitive Psychology*. (1999). DOI:https://doi.org/10.1002/(SICI)1099-0720(199911)13:1+<S73::AID-ACP631>3.3.CO;2-W.
- [2] Hu, Y. et al. 2008. Collaborative Filtering for Implicit Feedback Datasets. *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining* (Washington, DC, USA, 2008), 263–272.
- [3] Lops, P. et al. 2011. Content-based Recommender Systems: State of the Art and Trends. *Recommender Systems Handbook*. F. Ricci et al., eds. Springer US. 73–105.
- [4] MacLin, O.H. et al. 2005. PC\_eyewitness and the sequential superiority effect: Computer-based lineup administration. *Law and Human Behavior*. (2005). DOI:https://doi.org/10.1007/s10979-005-3319-5.
- [5] Mansour, J.K. et al. 2017. Evaluating lineup fairness: Variations across methods and measures. *Law and Human Behavior*. (2017). DOI:https://doi.org/10.1037/lhb0000203.
- [6] Meissner, C.A. and Brigham, J.C. 2001. Thirty Years of Investigating the Own-Race Bias in Memory for Faces: A Meta-Analytic Review. *Psychology, Public Policy, and Law*.
- [7] Parkhi, O.M. et al. 2015. Deep Face Recognition. *Proceedings of the British Machine Vision Conference 2015* (2015).
- [8] Peška, L. and Trojanová, H. 2017. Towards recommender systems for police photo lineup. *ACM International Conference Proceeding Series* (2017).
- [9] Shapiro, P.N. and Penrod, S. 1986. Meta-Analysis of Facial Identification Studies. *Psychological Bulletin*.
- [10] Steblay, N. et al. 2003. Eyewitness Accuracy Rates in Police Showup and Lineup Presentations: A Meta-Analytic Comparison. *Law and Human Behavior* (2003).
- [11] Valentine, T.R. et al. 2007. How can psychological science enhance the effectiveness of identification procedures? An international comparison. *Public Interest Law Reporter*. (2007). DOI:https://doi.org/10.1017/CBO9781107415324.004.