images by estimating the visual space covered by the images [3]. Sharma et al. [17] performs image summarization using SIFT features and topic modeling. This study is mainly aimed to summarize hundreds of photos taken by a single user.

III. Used Method

In order to analyze and choose the most representative images reflecting the events and issues of refugees in Turkey, we looked for the mostly shared visual elements linked to the trending topics determined from text analysis. The process to reach that objective includes following steps:

- Collecting relevant data
- Textual analysis to determine the trending events and issues.
- Visual analysis to define the representative image.

A. Collecting relevant data

To retrieve information from refugee related social media accounts, we analyzed public Twitter activity using Twitter API[1]. applying the method used in [4] consists of the following sub steps: Firstly, we tried to figure out a method to define the accounts of Syrian refugees in Turkey. Secondly, we traced back those chosen users' accounts, collecting the tweets. Finally we filtered out the irrelevant and unreliable data.

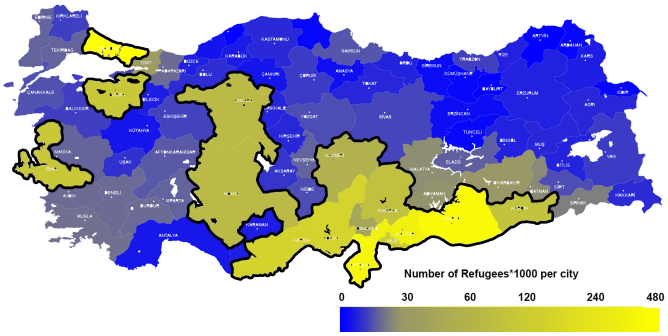

Fig. 1. Distribution of Syrian refugees in Turkish cities generated according to the statistics in [7].

Determining the accounts. To determine the refugee owned accounts we check the accounts associated with Arabic language, since refugees (95% [21]) have Arabic as their mother tongue and refugees are forming 90% of the registered foreigners in Turkey, Thus we assume that Twitter accounts in Turkey which uses Arabic as the main language are probably refugees'.

Then, we checked number of Syrian refugees registered in each Turkish city. Figure 1 shows the distribution of Syrian refugees in Turkey. The image is generated according to the statistics in [7] Using a logarithmic color code. The bold border shows the regions chosen in our study, which host the majority of Syrian refugees. The cities in these regions accommodate 90% of them while they only have 48% of the total Turkish population. Table I shows the number of refugees in chosen cities and the ratio of refugees to the city populations.

[1]https://developer.twitter.com/en/docs

TABLE I
Main Cities Hosting Refugees

City	N.Refugs	Ratio	City	N.Refugs	Ratio
Adana	150790	6.85%	Ankara	73198	1.37%
Bursa	106000	3.68%	G.Antep	329670	16.70%
Hatay	384024	24.69%	Istanbul	479555	3.24%
Izmir	108306	2.58%	K.Maras	90100	8.11%
Kayseri	59938	4.34%	Kilis	124000	95.15%
Konyai	73445	3.40%	Mardin	94340	11.85%
Mersin	146931	8.28%	Osmaniye	43773	8.38%
S.Urfa	420532	21.67%	Total	2685669	6.28%

Twitter API allows searching for recent (seven days) tweets according to keywords, location, and used language. To avoid a biased collection of data, we don't provide any keywords. We search for Arabic tweets in specific locations wherein refugees are accommodated intensely.

For practical purposes we limited the search to 1000 tweets per region. Later, we extracted the individual user IDs posting these tweets. As one account can post multiple tweets, we have less amount of user accounts then the number of tweets. As a result we collected a total of 5707 twitter users who were active recently. Table II shows the number of discovered users in each region.

TABLE II
Data collected: users- Tweets per city

City	N.Users	N.Tweets	City	N.Users	N.Tweet
Adana	160	127031	Ankara	1050	2297028
Bursa	2023	4543793	G.Antep	535	1139878
Hatay	12	13874	Istanbul	1183	2910948
Izmir	78	107263	K.Maras	35	13538
Kayseri	62	55275	Kilis	3	558
Konyai	137	158845	Mardin	30	24130
Mersin	269	461598	Osmaniye	22	11811
S.Urfa	108	156736	Total	5707	12022306

Collecting tweets data. Twitter API lets accessing up to last 3200 of a user's tweets including retweets. Therefore, we gathered tweets of each user we extracted in the previous step until the limit is reached or there is no more tweet from that user. In one query, it is possible to access only 200 tweets, thus, we run multiple queries to collect the maximum possible number of tweets. Table II includes the number of tweets collected from each city.

Data filtering. Firstly we have excluded the tweets from 2018, only tweets up to the end of 2017 were analyzed, since in 2018, millions of tweets (3,678,739) were retrieved in less than 20 days which are unreliable, due to the fact that tweeting 3200 tweets in those days means more than 160 tweets per day, implying that these tweets are neither coming from a real user, nor expressing an individual user opinion. And among the traced accounts, there are ones which do not belong to individual users but to press or companies for instance. These accounts mostly post with a very high frequency including a big ratio of retweets. This information helps differentiating the